D2.4

Classification of HILP events – a Machine Learning approach

Date

30.03.2026

Authors: Arka Bhattacharyya, Nazli Yonca Aydin, Martijn Warnier

Funded by the European Union

Project funded by the European Union's Horizon Europe and the UK Research and Innovation (UKRI) programme under the UK government's Horizon Europe funding guarantee grant agreement n°10062626

D2.4

Classification of HILP events – a Machine Learning approach

Revision 6.0

Grant Agreement	101121356
UKRI numbers	10062626
Call identifier	HORIZON-CL3-2022-DRS-01
Project full name	AGnostic risk management for high Impact Low probability Events
Due Date	31/03/2026
Submission date	31/03/2026
Project start and end	01.10.2023 - 30.09.2027
Authors	Arka Bhattacharyya, Nazli Aydin, Martijn Warnier

Abstract

In AGILE project's task T2.3, a two-stage machine learning framework was applied to a historical database of over 6000 disaster records from the 21st century. In the first stage, unsupervised learning via the K-means algorithm successfully classified nearly 9% of historical disasters as HILP events thus providing a clear differentiation from standard incidents based on historical impact patterns. Classifying these incidents led to the identification of their characteristics based on empirical evidence. The second stage utilized a supervised Random Forest classification model to identify the underlying drivers of these historical HILP events. The analysis revealed that the proportion of the population affected per million representing population exposure and hazard type are the two most influential predictors of HILP disasters. Furthermore, socio-economic factors specifically institutional trust, community engagement, and unemployment were identified as critical determinants of HILP disasters' devastation potentials. Findings were validated through internal stakeholder reviews in Rotterdam and external peer feedback at the International Climate Resilience Conference in Munich (2025). Additionally, we conducted an online stakeholder survey within AGILE's network where majority of the survey respondents agreed with our methodology and its outcomes. The findings of the machine learning models led to the development of actionable mitigation measures against future HILP events.

Document revision history

Issue	Date	Comment	Author
V0.1	12.01.2026	Report Structure	Arka Bhattacharyya
V1.0	29.01.2026	First Draft	Arka Bhattacharyya
V1.1	11.02.2026	Review of First Draft	Nazli Yonca Aydin
V2.0	16.02.2026	Second Draft	Arka Bhattacharyya
V3.0	02.03.2026	Third Draft (with Validation)	Arka Bhattacharyya

V3.1	03.03.2026	Review of Third Draft	Martijn Warnier
V3.2	04.03.2026	Review of Third Draft	Nazli Yonca Aydin
V4.0	06.03.2026	Fourth Draft	Arka Bhattacharyya
V4.1	12.03.2026	Review of Fourth Draft	Rabea Schulz
V4.2	15.03.2026	Review of Fourth Draft	Davide Ferrario
V5.0	18.03.2026	Fifth Draft	Arka Bhattacharyya, Nazli Yonca Aydin, Martijn Warnier
V5.1	23.03.2026	Review of Fourth Draft	Rabea Schulz
V5.2	23.03.2026	Review of Fourth Draft	Saman Ghaffarian
V5.3	23.03.2026	Review of Fourth Draft	Gordana Cveljo
V5.4	23.03.2026	Review of Fourth Draft	Davide Ferrario
V6.0	27.03.2026	Final Draft	Arka Bhattacharyya

Acknowledgment

Project funded by the European Union's Horizon Europe under the grant agreement n°101121356 and the UK Research and Innovation (UKRI) programme. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Nature of the deliverable[1]		R

Dissemination level

PU	Public, fully open. e.g., website	✓
SEN	Sensitive, limited under the conditions of the Grant Agreement	
CL	Classified information under the Commission Decision No2015/444	

[1] Deliverable types:

R: document, report (excluding periodic and final reports).

DEM: demonstrator, pilot, prototype, plan designs.

DEC: websites, patent filings, press and media actions, videos, etc.

OTHER: software, technical diagrams, etc.

Copyright notice

Table of Contents

List of Figures

List of Tables

Abbreviations

Tx.x	Task x.x
HILP	High Impact Low Probability
WP	Work Package
PCA	Principal Component Analysis
SOM	Self Organizing Map

Dx.x	Deliverable x.x
DoA	Description of the Action
OECD	Organisation for Economic Co-operation and Development
GDP	Gross Domestic Product
ID	Identifier
CPI	Consumer Price Index
WB	World Bank
HDI	Human Development Index
CSPI	Civil Society Participation Index
LM	Lower-middle (income)
UM	Upper-middle (income)
L	Low (income)
H	High (income)
WCSS	Within-Cluster Sum of Squares
OWID	Our World in Data
PDP	Partial Dependency Plots
SMOTE	Synthetic Minority Oversampling Technique

Executive Summary

The AGILE project's Task 2.3 (T2.3) focuses on the analysis of High Impact Low Probability (HILP) events through a comprehensive machine learning framework. For this task, we developed a historical disaster database containing over 6000 disaster records from the 21st century to support this research. The primary objective involved the identification and classification of HILP events using machine learning methods. A two-stage analytical approach was successfully implemented to categorize disasters and uncover their underlying drivers. The K-means clustering algorithm found 2 clusters of disasters based on fatalities, economic losses, and precedence. The clusters were utilized to distinctly separate HILP disasters from standard events. Approximately 9% of the historical disasters belonging to one cluster that showed statistically significant higher level of fatalities and economic losses but lacked a precedence were labeled as HILP events. Supervised machine learning was subsequently applied to link these classifications to specific risk factors. The Random Forest algorithm was used to identify the most influential predictors of disaster escalation. The proportion of the population affected per million was confirmed as the dominant driver for HILP outcomes. Socio-economic factors, including institutional trust and community engagement, were found to be critical determinants of systemic failure. The findings of the machine learning models led to the development of actionable guidelines and threat-agnostic mitigation measures against HILP events. Validation of the research findings was conducted through engagement with internal and external stakeholders. Model outcomes were presented and refined at the AGILE General Assembly in Rotterdam in September 2025 and the International Climate Resilience Conference in Munich in October 2025. Further, we validated the outcomes of the machine learning models through an internal stakeholder survey within the project's network, where the majority of the respondents agreed with our methodology and outcomes. Thus, the 3 objectives identified within T2.3 were successfully completed within the planned project timeline.

Chapter 1 Introduction

The AGILE project is divided into 8 work packages (WPs). The second work package, i.e., WP2, focuses on developing a database of historical High Impact Low Probability (HILP) events (T2.2). The HILP events are typically characterized by limited foresight, long recurrence periods, and extreme potential for devastation [1]. The HILP database facilitates analysis of HILP events to understand their underlying patterns so that appropriate risk management interventions can be planned. The task T2.3 within WP2 is about this analysis of historical HILP events using machine learning models. It has 3 main objectives. They are as follows:

- **1. Identify and classify HILP events.**

Use unsupervised Machine Learning methods like Principal Component Analysis (PCA) or Self Organizing Map (SOM) to detect patterns in global disaster datasets. Categorize events into meaningful clusters to understand common drivers and characteristics.

- **2. Analyze factors influencing escalation and variation leading to developing actionable guidelines.**

Examine the social, spatial, cultural, and contextual factors that differentiate or intensify HILP events. Identify both similarities and dissimilarities to uncover what triggers event severity. Use these insights to create guidelines and recommend tailored mitigation measures for specific event types and contexts.

- **3. Validate findings.**

Collaborate with stakeholders within AGILE's network to verify model outcomes and ensure practical relevance.

The main focus in T2.3 is to utilize machine learning models. It requires building a large dataset that can be analyzed using machine learning models. Hence, we developed a large dataset of historical disaster events of this century from publicly available data sources. The dataset contained information about more than 6000 disaster events between 2000 and 2023. We further complemented the disaster data with relevant exposure, vulnerabilities, and coping capacity variables that we identified in task T2.1 of WP2, which has already been reported in D2.1.

For the first objective, we utilized unsupervised machine learning models. As stated in the Description of the Action (DoA), we first tested Self Organizing Maps (SOMs) to check if the algorithm can separate HILP and non-HILP events. SOMs are typically valued for their ability to project high-dimensional data onto a two-dimensional grid while preserving the topological structure of the dataset. Clear boundaries between the two event types were not established by the algorithm. Due to the complex internal structure of the SOMs, explainability of the results was significantly limited. This lack of explainability was

identified as a major drawback. Usefulness for practitioners was deemed limited because the maps were difficult to interpret. In contrast, a more transparent method was required to ensure the results could be easily understood and applied by stakeholders.

The K-means clustering algorithm was subsequently adopted as a more effective and explainable alternative. A centroid-based partitioning approach was utilized to group the data based on direct mathematical proximity. This method provided a more intuitive and transparent separation between clusters. Distinct separation between HILP disasters and non-HILP disasters was successfully achieved through this approach. Based on this separation, we identified several characteristics of HILP disasters based on historical evidence.

These data-driven classifications were then used as target variables in supervised machine learning models. This transition ensured that the research outcomes remained interpretable and actionable for the project partners. The supervised machine learning model is designed to assess to what extent key dimensions of disaster risk can explain the separation between HILP and non-HILP events, and to identify which combinations of factors most strongly drive HILP outcomes. Therefore, it relates to objective 2, Analyze factors influencing escalation and variation, within task T2.3. Model performance is evaluated using standard classification metrics that distinguish correctly and incorrectly classified HILP and non-HILP events, thereby providing a rigorous test of explanatory power. The outcomes of the supervised machine learning model led to the development of hazard agnostic mitigation measures against future HILP disasters. Taken together, this two-stage approach first uncovers HILP structures directly from observed disaster impacts and then links these structures back to underlying risk drivers.

To achieve the third objective, i.e., validate findings and develop actionable guidelines, the outcomes of this 2-stage analysis were presented to both internal and external stakeholders. For internal validation, we presented the research outcomes at AGILE project's 4th General Assembly in September 2025 in Rotterdam, Netherlands. We incorporated suggestions from the stakeholder partners in refining the machine learning models. For external validation, we presented the research outcomes at the International Climate Resilience Conference in October 2025 in Munich, Germany. We received some feedback from the audience, which further helped us refine the models. Lastly, we conducted an internal stakeholder survey within AGILE's network to validate the data-driven methodology and its outcomes to prove their robustness. By implementing unsupervised machine learning followed by supervised machine learning and presenting outcomes to both internal and external stakeholders, task T2.3 has been completed successfully. This report summarizes how the activities in T2.3 were planned, executed, and the findings from the activities.

Chapter 2 Methodology

Figure 1 shows the methodological flowchart for task T2.3. The flowchart presents a four-step methodology for identifying and explaining High-Impact Low-Probability (HILP) disasters using Machine Learning.

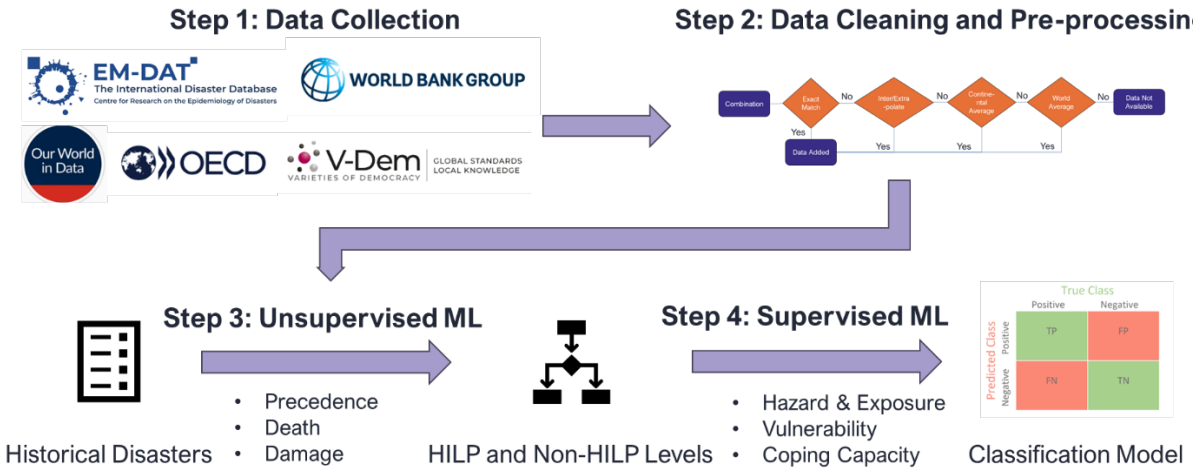


Figure 1 Methodological Flowchart for Task T2.3

The process starts with data collection from multiple global databases. These sources include EM-DAT, the World Bank, Our World in Data (OWID), Organization for Economic Co-operation and Development (OECD), and V-Dem. Details can be found in Table 1. Each source provides different but complementary indicators on disasters, society, and governance. In deliverable D2.1, we set an ambition for this data collection based on the outcomes of a stakeholder survey and requirement workshop. Based on that, we collected data for several indicators belonging to hazard and exposure, vulnerability, and coping capacity dimensions of disaster risk [2]. These variables have been explained in detail in the following sections. Together they create a broad and rich dataset on historical disasters and their contexts.

Table 1 Collected Data and Their Sources

No.	Variable	Data Granularity	Data Source
1	Disaster ID	Disaster specific	EM-DAT
2	Country	Disaster specific	EM-DAT
3	Continent	Disaster specific	EM-DAT
4	Year	Disaster specific	EM-DAT
5	Hazard	Disaster specific	EM-DAT
6	No. of deaths	Disaster specific	EM-DAT
7	No. of people affected	Disaster specific	EM-DAT

No.	Variable	Data Granularity	Data Source
8	Economic losses in nominal terms	Disaster specific	EM-DAT
9	Economic losses CPI adjusted	Disaster specific	EM-DAT
10	Population of country	Country specific	WB
11	Nominal GDP of country	Country specific	WB
12	Human development index (HDI)	Country specific	WB
13	Per capita real GDP	Country specific	WB
14	Share of population in poverty	Country specific	WB
15	Literacy rate	Country specific	WB
16	Unemployment rate	Country specific	WB
17	Percentage of population over 65 years	Country specific	WB
18	Percentage of population between 0 and 14 years	Country specific	WB
19	Percentage of population with electricity access	Country specific	WB
20	Fixed telephone subscription per 100 people	Country specific	WB
21	Percentage of population using internet	Country specific	WB
22	Percentage of population using safely managed drinking water services	Country specific	WB
23	Percentage of population using safely managed sanitation services	Country specific	WB
24	Hospital beds per 1000 people	Country specific	WB
25	Prevalence of food insecurity	Country specific	WB
26	Prevalence of undernourishment	Country specific	WB
27	Government effectiveness	Country specific	WB
28	Percentage of population trusting police, armed forces, civil services, government, judiciary	Country specific	Integrated Values Surveys with major processing by Our World in Data
29	Civil society participation index (CSPI)	Country specific	V-Dem
30	Foreign aid received by countries	Country specific	OECD

In the second step, the data is cleaned and pre-processed. Variables from different sources are combined into a coherent structure. Missing information is treated using clear decision rules. Some variables are imputed (labeled), others are averaged or normalized, and some are dropped if they are not usable. This step ensures consistent scales, units, and formats across all indicators. It prepares the dataset so that machine learning models can work reliably.

The third step applies unsupervised machine learning to the cleaned historical disaster data. The focus is on outcome variables such as number of deaths, and economic damage as well as historical precedence. These variables capture the severity and impact of past events. Using clustering or related unsupervised techniques, disasters are grouped into patterns without predefined labels. We investigated the clusters to see whether disasters can be separated into HILP and non-HILP categories from impact patterns, using a purely data-driven method. Since HILP disasters record unprecedented high levels of damage, we hypothesized that unsupervised machine learning algorithms would be able to identify these events and create a separate cluster of them.

The fourth step introduces supervised machine learning to explain these categories. The goal of this step is to see if the dimensions of disaster risk, i.e., hazard and exposure, vulnerability, and lack of coping capacity [2] can predict these HILP and non-HILP labels that we derived from step three. We also wanted to identify the most influential predictors among the dimensions of disaster risk and investigate how they influence the HILP and non-HILP separation from this analysis. The HILP and non-HILP labels from step three are utilized as the target variable. Explanatory variables representing dimensions of disaster risk, including hazard and exposure, vulnerability, and coping capacity were used to explore their predictive capabilities using Random Forest classification model.

To summarize, unsupervised machine learning algorithm used fatalities, damages, and precedence to separate historical disasters into different clusters that led to labeling them into HILPs and non-HILPs whereas supervised machine learning used variables related to dimensions of disaster risk to predict those labels to understand which predictors are more influential and how they influence the HILP/non-HILP separation.

To train the Random Forest classification model, the labelled disaster dataset (originating from unsupervised machine learning model) is divided into training and testing subsets ensuring model generalizability. A common split of 80 percent for training and 20 percent for testing is applied to the historical records. The supervised model is trained on the first subset to learn patterns between risk drivers and disaster categories. Predictions are then generated for each individual disaster event within the unseen testing set.

Model performance is evaluated by comparing these predictions against the known labels in the test set that were derived from the unsupervised machine learning model. A confusion matrix is used to distinguish true positives, false positives, true negatives, and

false negatives. This evaluation assesses how well the risk dimensions explain the separation of specific HILP and non-HILP events.

Overall, the flowchart in Figure 1 shows a transparent pipeline from raw data to actionable insights. Data from diverse sources are harmonized, clustered into impact-based levels, and then used to build a classification model to identify the influential predictors and their influence on the target variable. The method first discovers HILP structures in the data and then tests which risk factors drive these structures. This combination of unsupervised and supervised approaches supports both discovery and explanation of rare, high-impact disasters.

2.1. Data Collection

As part of task T2.1, we identified the content requirements for the HILP reference library from relevant stakeholders through a survey and a workshop. The requirements set our ambition for the data collection to build the database. This was reported in deliverable D2.1

As mentioned earlier, we built a disaster database for machine learning-based analysis of HILP events. The dataset contained disaster information for more than 6000 disasters that occurred between the year 2000 and 2023. We limited ourselves to this period due to limited data availability of the drivers of disaster risk, i.e., exposure, vulnerability, and lack of coping capacity before this century. Furthermore, we only focused on disasters caused by earthquakes, storms, floods, and droughts as these 4 hazards collectively accounted for approximately 85% of the disasters between 2000 and 2019 [3]. As there is no dedicated database for historical HILP events, we took this approach, where we worked with a larger disaster dataset and let the unsupervised machine learning model separate HILP and non-HILP events into different clusters.

For each disaster event, we collected several variables that are listed in Table 1. Disaster-specific information, including unique identifiers, geographical locations, hazard types, and temporal data, was integrated into the initial dataset. Impact metrics such as fatalities, the number of people affected, and economic damage were also recorded. These disaster-specific variables were utilized by unsupervised machine learning models to derive the HILP and non-HILP labels.

For the supervised machine learning stage, a secondary set of predictors was required to explain these classifications. Due to limitations in localized data availability, country-level indicators were adopted as representative proxies for the disaster zones. For example, real GDP per capita was utilized at the national level because specific economic data for the affected local areas were unavailable. This approach ensured that a consistent set of

socio-economic and vulnerability metrics could be applied across the entire global database.

The use of national-level indicators as proxies for localized disaster zones is a significant limitation. Sub-national variations in vulnerability and resilience are often masked by these broad averages. Disaster outcomes are strongly influenced by the specific local geography and local contexts. Important predictive nuances are lost when uniform national values are applied to localized events. Finer spatial resolution was not utilized due to the absence of disaster polygons in global databases like EM-DAT. This data constraint prevents the precise mapping of risk drivers for thousands of historical events.

Event-level data were primarily sourced from EM-DAT. For every disaster, the database records a unique disaster identifier, country, continent, year, and hazard type. It also includes the number of deaths, the number of people affected, and economic losses in nominal and CPI-adjusted terms. These variables describe the direct human and economic consequences of each event and support the quantification of disaster severity and precedence.

To contextualize the events, disaster records are linked to **country-level indicators** from the World Bank. Demographic variables include total population, age structure (shares over 65 years and between 0 and 14 years), and literacy rate. Economic conditions are represented by real GDP per capita, poverty rates, and unemployment rates. Infrastructure and service access are captured by electricity access, safely managed water and sanitation, hospital beds per 1000 people, fixed telephone subscriptions, and internet usage.

The database also integrates indicators of food security, health, and governance. Food insecurity and undernourishment measure chronic vulnerability that can shape disaster impacts. Governance quality is represented by the government effectiveness index, which reflects public service quality and policy implementation [4]. Institutional trust covers confidence in the police, armed forces, civil service, government, and judiciary, based on survey data processed by Our World in Data [5]. These indicators support the analysis of how public institutions relate to HILP outcomes.

Finally, the dataset includes variables on civil society and international support. Civil society engagement is measured by a civil society participation index from V-Dem [6]. Foreign aid received by each country, obtained from the OECD, represents external financial assistance. These indicators provide information on participation, support, and potential capacity for preparedness and recovery. Together, the country-level variables describe hazard exposure, vulnerability, coping capacity, and governance environments, enabling a comprehensive and data-driven exploration of conditions under which disasters escalate into HILP events.

2.2. Data Cleaning and Pre-processing

The data imputation scheme specifying how missing values in country-level vulnerability and coping capacity variables were handled before the machine learning analysis is shown in Figure 2. The scheme operates at the level of unique disaster–country–year combinations, which are defined using hazard type, country ISO code, year, and continent. For each combination, the goal is to attach a complete set of country-level indicators to the HILP database.

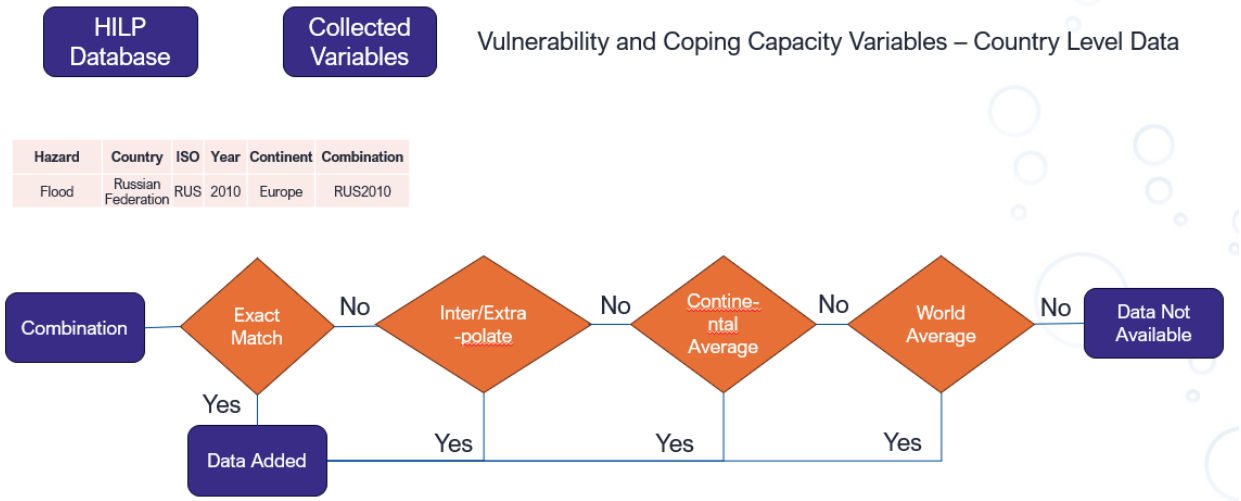


Figure 2 Data Imputation Scheme

The imputation process has a number of steps as shown in Figure 2. **Step 1: Exact Match Identification** An exact match between the disaster event and the collected variables is first sought. The value is directly added to the database if an exact country-year observation is found. No further imputation is required for these entries. The record is then flagged as successfully populated within the system.

Step 2: Temporal Interpolation and Extrapolation Neighboring years for the same country are searched when an exact match is absent. Linear interpolation is applied if data is available both before and after the target event year. Extrapolation is used when data exists only for preceding or succeeding years. These computed values are then integrated into the dataset to maintain temporal consistency.

Step 3: Continental Average Substitution A continental average is calculated if country-specific temporal data is unavailable. Averages are derived from all countries within the same continent for that specific year. This proxy is used to represent the missing value while maintaining broad regional characteristics. The disaster combination is then updated with this regional average to bridge the data gap.

Step 4: Global Average Substitution A global average is computed as a final estimation step if regional data is also missing. The world average is calculated using data from all available countries for that specific year. This global value is assigned to ensure the

variable remains defined for the record. The continuity of the dataset is maintained through this hierarchical fallback approach.

Step 5: Data Point Finalization The entry is officially marked as not available if no value is yielded by any of the previous steps. The variable remains missing for that specific disaster-country-year combination.

Each step in the decision path is documented to provide transparency regarding data quality. This structured scheme is used to maximize data coverage while preserving consistency for later sensitivity analyses.

Variables from different sources had varying scales, units, and distributions. All variables were checked for consistent data types. Categorical variables such as hazard type and continent were encoded appropriately. Numerical variables were converted to uniform formats. Continuous variables were normalized to ensure comparability across indicators. This step prevented variables with large numeric ranges from dominating the machine learning models. It also facilitated the interpretation of variable contributions in clustering and classification tasks.

The final dataset had 6211 data points. Figure 3 provides distinct insights into hazard types, economic development contexts, geographic concentration, and temporal patterns, collectively supporting the data exploration phase prior to machine learning analysis of HILP events.

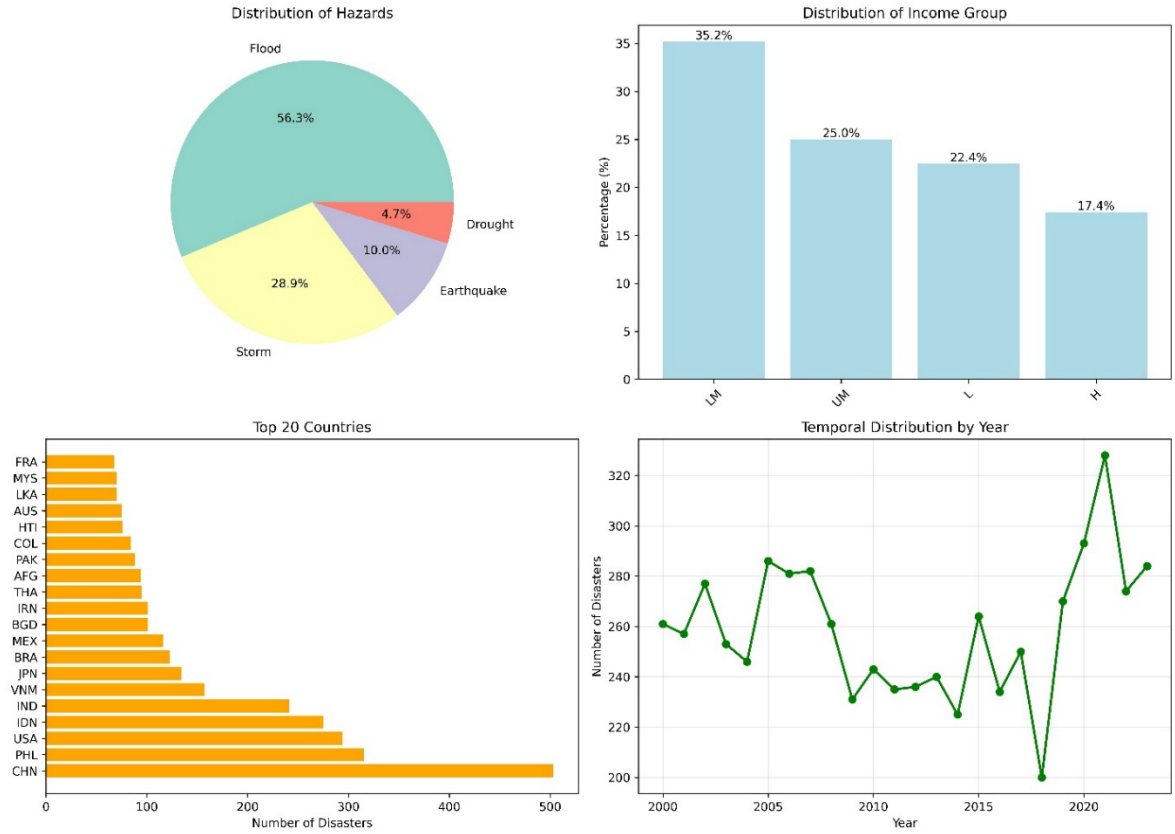


Figure 3 Final Data and Insights into Hazard Types, Economic Development Contexts, Geographic Concentration, and Temporal Patterns

The top-left pie chart illustrates the percentage distribution of disasters by hazard category. Floods dominate the database with 56% of recorded events, reflecting their prevalence across diverse geographic and climatic contexts. Storms follow at 29%, while earthquakes (10%) and droughts (5%) constitute smaller shares, consistent with the dataset's focus on these four major hazards that collectively account for 85% of disasters between 2000 and 2019 [3]. This distribution highlights floods as the primary analytical focus for identifying HILP patterns due to their numerical dominance.

The top-right bar chart shows the distribution of disasters across income groups. Lower-middle-income (LM) countries experience the highest share of disasters at 35%, indicating disproportionate exposure and vulnerability at this development stage. Upper-middle-income (UM) countries follow at 25%, suggesting a significant decline as countries advance economically. Low-income (L) countries account for 22%, which is notable given their presumed higher vulnerability compared to UM countries, possibly reflecting smaller populations or lower hazard exposure in these regions. High-income (H) countries represent the smallest share at 17%, demonstrating that disaster frequency decreases substantially with economic development.

The bottom-left horizontal bar chart reveals the spatial concentration of disasters, ranking the top 20 countries by event count. China leads with approximately 550 events, followed by the Philippines, the United States of America, Indonesia, and India. This concentration in densely populated, hazard-prone nations underscores the role of exposure in driving disaster frequency. The geographic focus on Asia and the Americas (among the top 20 countries) provides context for interpreting vulnerability and coping capacity indicators at the country level.

Finally, the bottom-right line chart depicts the temporal evolution of disaster counts by year. The dataset begins at approximately 260 disasters in 2000. Disaster counts fluctuate considerably throughout the two-decade period, ranging between 200 and 330 events annually. The lowest point occurs in 2017 with roughly 200 disasters recorded. The highest peak is observed in 2021 with approximately 330 disasters, representing a dramatic surge in recorded events.

2.3. Unsupervised Machine Learning Models

Unsupervised machine learning models were adopted to see if the algorithms can separate HILP disasters from the non-HILP ones, which is essentially objective 1, i.e., identify and classify HILP events, of task T2.3. Before executing the unsupervised machine learning algorithm, four steps were implemented.

Step 1: Impact Variable Normalization: HILPs, i.e., high impact low probability disasters have two dimensions, high impact and low likelihood. For impact, we used disaster-related

deaths and economic losses [3]. But before using them in the unsupervised machine learning models, we first normalized the values for deaths and economic losses with respect to the total population of the affected country and its nominal GDP, respectively. This was done to neutralize the effects of different population size and size of economy. So, the impact variables used in unsupervised machine models were deaths per million population and economic losses per billion GDP.

Step 2: Precedence and Likelihood Categorization: For likelihood, the analysis introduced a new binary variable to capture disaster precedence at the country level. This variable indicates whether a country had already experienced a more severe event of the same hazard type before the focal disaster. As mentioned earlier, we had four hazard types in our dataset. They are floods, storms, earthquakes, and droughts. Now the intensity or magnitude of these hazards are expressed differently. For instance, earthquakes' intensities are often expressed in Richter scale. On contrary, the intensities of storms are often expressed in wind speed and rainfall amount. Due to this difference, we adopted the number of people affected by a disaster as a representative variable for disaster severity. This variable expressed the severity of all hazard types uniformly for all hazards. A value of 1 is assigned if at least one prior event from the same hazard caused a higher number of people affected than the event under study. Otherwise, the variable takes the value 0. The construction of this indicator relies on recorded disaster histories from the EM-DAT database, which provides consistent information on past events and their impacts. The precedence variable is grounded in the idea that past disasters create opportunities for learning, foresight, and improved preparedness [7]. Historical experience may lead to better risk awareness, stronger institutions, and more effective early warning systems. It may also stimulate investments in protective infrastructure and emergency response capacity. We found a statistically significant effect of precedence on deaths and damages due to disasters. It can be found in Appendix A.

Step 3: Feature Scaling for Machine Learning: These three variables, i.e., precedence, deaths per million population, and economic losses per billion GDP were used in unsupervised machine learning models to investigate whether the algorithm can separate HILPs and non-HILPs. But before doing that, all three variables were normalized through a min-max normalization to eliminate the differential effect of three variables due to their differences in scale.

Step 4: Integration into the Unsupervised Framework: The final dataset, consisting of scaled precedence and impact metrics, was fed into the unsupervised clustering models. The objective of this step is to investigate whether the algorithm can naturally group these historical disaster events into different clusters.

Step 5: Deriving HILP and non-HILP Labels: It has been explained earlier that we hypothesized the unsupervised machine learning algorithm would be able to create

separate clusters of historical disasters that will lead to the identification of historical HILP events. To put HILP labels on them, we analyzed each cluster based on the impact and precedence variables. We assume this analysis will lead to putting labels on disasters belonging to one or more clusters as HILPs and the rest as non-HILPs.

Self-Organizing Maps

Self-organizing maps are a type of neural networks that rearrange themselves to reflect patterns in data [8, 9]. They take complex, high-dimensional data and lay it out on a simple two-dimensional grid. Each position on the grid acts like a prototype data point. Nearby positions on the grid represent data points that are similar to each other. Over many training steps, the map learns to place similar cases close together and different cases far apart.

Training works through competition and cooperation. For each data point, the map finds the grid cell whose weights are most similar, i.e., the winner or best matching unit. The winner and its neighbors then adjust their weights slightly toward that data point. Repeating this for many data points makes the map gradually organize itself. The result is a visual layout that clusters similar observations. Thus, the neighborhood structure is preserved, meaning that similar data points are mapped to adjacent cells. Theoretically, this makes patterns, groups, and anomalies in the data easier to see and interpret. However, in practice, the outcomes are not always easy to interpret.

K-means Clustering

K-means clustering groups data points into a fixed number of clusters [10]. You first choose how many clusters, i.e., K are required. The algorithm starts by picking K random points as initial centroids. Each data point is then assigned to the nearest centroid based on distance.

After assignment, the centroids are updated to the average position of all points in their cluster. This process is repeated. Points get reassigned to the closest centroid. Centroids move to new average positions. The algorithm stops when assignments no longer change or after a set number of iterations. The results show distinct groups, where points within each cluster are more similar to each other than to points in other clusters.

K-means clustering requires choosing the number of clusters, K, beforehand. The elbow method helps to find the best K [11]. The method runs K-means for different values of K, typically from 1 to 10 or 15. For each K, the Within-Cluster Sum of Squares (WCSS) is calculated. WCSS measures how spread-out points are from their cluster center. As K increases, WCSS always decreases as more clusters fit the data better. Initially, WCSS drops sharply with each new cluster. After an optimal point, the decrease slows down significantly. This creates a bend or elbow shape in the plot. The elbow point marks where adding more clusters gives little improvement. The K at this elbow is considered as the

optimal number of clusters. The method balances underfitting (too few clusters) and overfitting (too many clusters).

2.4. Supervised Machine Learning Models

The unsupervised machine learning models yielded HILP and non-HILP labels for the datapoints. The objective of the supervised machine learning model is to explore whether the dimensions of disaster risk, i.e., hazard and exposure, vulnerability, and lack of coping capacity [2] can explain the HILP and non-HILP labels and, among these dimensions, which is more influential in explaining that label. This part of the analysis essentially addressed the second objective, i.e., analyzing factors influencing escalation and variation, of task T2.3 within WP2. Supervised machine learning models have two types of variables. First, the target variable, which is HILP or non-HILP label of a disaster derived from unsupervised machine learning model. Next, explanatory variables or predictors are the variables corresponding to the dimensions of disaster risk that are used in the supervised machine learning model to predict the target variable. They are as follows.

Hazard and Exposure

It has been explained before that the dataset contained four *hazard* types, which are floods, storms, earthquakes, and droughts. So, in the analysis, *hazard* has been used as a categorical variable. For *exposure*, we used the number of people affected per million population as a representative variable. This was done due to lack of publicly available global disaster specific exposure datasets since the turn of the century.

Vulnerability

Vulnerability captures the susceptibility of populations and systems to disaster impacts [12]. It includes human development, economic, demographic, and nutritional indicators. These variables reflect pre-existing conditions that amplify disaster consequences.

The *Human Development Index* (*HDI*) measures overall development through life expectancy, education, and income [13]. Higher *HDI* suggests better baseline conditions for disaster response. *Per capita real GDP* indicates economic resources available for preparedness and recovery. Lower real GDP per capita correlates with higher vulnerability. The *poverty rate* measures the proportion of the population below the national poverty line. Poverty limits access to protective assets and services.

Literacy rate reflects educational attainment across the population. Higher literacy supports better understanding of warnings and instructions. *Unemployment rates* indicate economic instability and dependence on vulnerable livelihoods. High unemployment reduces household resilience. *Population over 65 years* captures the elderly share, who

face greater mobility and health limitations during disasters. *Population under 14 years* measures dependent children, who require protection and care during crises.

Food insecurity prevalence shows households lacking reliable access to adequate food. *Undernourishment* measures the share of population with insufficient caloric intake. Both nutritional indicators signal chronic weakness that exacerbates disaster mortality and morbidity. Together, these vulnerability indicators provide a comprehensive view of societal fragility across multiple dimensions.

Coping Capacity

Coping capacity reflects a society's ability to absorb, adapt to, and recover from disasters [14]. These indicators measure infrastructure access and institutional resilience. Higher coping capacity reduces the likelihood of HILP outcomes.

The first set of coping capacity variables reflects "*access to basic infrastructures*". Access to six infrastructures (electricity, telephone, internet, drinking water, sanitation, and hospital) was considered in this research. *Electricity* access measures the percentage of population with reliable power. *Telephone* subscriptions indicate communication infrastructure availability. *Internet* access reflects information dissemination capacity. *Safely managed drinking water* access shows water security during crises. *Safely managed sanitation* prevents secondary health risks. *Hospital beds per 1000 people* measure medical response capacity.

The next set of variables reflects the governance aspect of disaster risk. We considered four such variables. The first variable is a composite index developed by the World Bank named *government effectiveness*, which "captures perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies" [4]. *Trust in police, armed forces, civil services, government, and judiciary* averages the percentage of population that thrusts these organizations reflecting public confidence in emergency authorities [5]. *Civil Society Participation Index (CSPI)* measures the strength and engagement of civil society organizations in a country. It assesses how routinely policymakers consult major civil society groups [6, 15]. The index evaluates citizen involvement in these organizations. It also considers women's participation levels. Higher CSPI values indicate robust civil society that influences public policy. Strong civil society improves disaster preparedness through advocacy and community networks. *Foreign aid* received indicates external support for preparedness and recovery [16]. These institutional factors determine response coordination and social cohesion during disasters.

Continent

In addition to the dimensions of disaster risk, we considered *continent* as another categorical variable to take into account any continent specific characteristics such as higher frequency of natural hazards, etc., that we might have missed in the above explained variables. The variable *Continent* controls for regional climate patterns and tectonic activity differences. It also accounts for varying data quality and reporting standards across world regions.

Random Forest Classification

For supervised machine learning, we adopted the Random Forest classification model [17]. Random Forest Classification builds multiple decision trees to make predictions. Each tree votes on the final label, i.e., HILP or non-HILP label of a disaster event based on the predictors. The majority vote determines the outcome. This ensemble approach reduces overfitting compared to single trees. Trees train on random subsets of data and features. This randomness improves generalization. Also, the method handles both categorical and numerical disaster risk variables effectively. Hence, it was adopted for objective 2 within T2.3.

Decision trees split data based on feature thresholds. Random Forest selects a random sample of features at each split. Bootstrap sampling creates diverse training sets for each tree. In bootstrap sampling, subsets of the original disaster data are randomly selected with replacement to train each individual tree. This means some data points may be repeated in a single tree's training set, while others are omitted. This technique ensures that no two trees are identical, as they are exposed to different versions of the dataset. Variability is increased through this randomness to prevent the model from overfitting or simply memorizing specific data points. It provides class probabilities alongside predictions.

Out-of-bag samples validate performance during training. These samples consist of the data points that were excluded from the training set of a particular tree during the bootstrap process. The predictive accuracy of each tree is tested on these unseen records while the forest is still being built. An internal estimate of the model's error is provided through this process, which reduces the immediate reliance on a separate validation dataset.

Train Test Split

The dataset was split into training and test sets using an 80/20 stratified sampling method. Stratified sampling ensured similar distribution of HILP and non-HILP disasters among training and test set. Eighty percent of the data was allocated for model training in the training set. The remaining twenty percent was reserved for final performance evaluation on unseen data. Robust generalization was ensured through this separation.

Hyperparameter Tuning

Hyperparameter tuning optimizes model performance. The Random Forest algorithm has several hyperparameters that need to be tuned for optimal model performance. Key parameters include number of trees, maximum depth, and minimum samples per split. Here, we varied the number of trees between 100 and 500, maximum depth between 7 and 50, and minimum samples per split between 20 and 100.

For tuning, we used a grid search approach. Grid search systematically evaluates all parameter combinations. It creates a complete grid from specified ranges of hyperparameters. For example, 5 tree values × 8 depth values × 5 split values yield 200 total combinations. Each combination trains a full Random Forest model. To prevent the model from simply memorizing the training data, a technique called k-fold (k = 5) cross-validation is employed. The dataset is partitioned into 5 subsets or folds. The model is trained on 4 folds while the remaining fold is used for validation. This cycle is repeated 5 times so that every data point is used for both training and testing. From the cross-validation, the combination with highest average F1-score or accuracy is selected. Accuracy measures the overall proportion of correct predictions out of all predictions made. F1-score is the harmonic mean of precision and recall, providing a single metric that balances both when their values differ. Precision measures the proportion of true positives (i.e., HILP disasters predicted as HILPs) predictions among all positive predictions (i.e., sum of actual HILP disasters and predicted HILP disasters) made by the model. Recall measures the proportion of true positive instances correctly identified among all actual positive instances (i.e., actual HILP disasters). These four metrics are also used to quantify and measure the predictive performance of supervised classification models.

Variable Importance Analysis

Variable importance measures each predictor's contribution to predictions. Two variable importance analyses have been conducted.

First, Random Forest calculates importance using the mean decrease in impurity. Gini impurity reduction quantifies splits' effectiveness. Importance sums contributions across all trees. Gini impurity shows how pure or mixed a group of disasters is. Pure groups contain only HILP disasters or non-HILP disasters. Mixed groups have both types. A score of 0 means perfectly pure. A score of 0.5 means maximum mixture. When a tree splits data, the improvement in purity is measured. The difference between before-split-impurity and after-split-impurity is calculated. Bigger improvements mean better splits. The feature creating that split gets credit. This happens at every split throughout the entire tree. All improvements from one feature are added together. The same process runs separately in every tree of the forest. Final scores average the improvements across all trees. Features

that consistently create the best splits rank the highest. This method is known to be biased toward continuous variables or features with many unique values. Hence, another type of variable importance analysis was performed.

Permutation-based variable importance is evaluated by measuring the decrease in model performance when a feature's values are randomly shuffled. This shuffling process is repeated multiple times to ensure the stability of the results. A significant drop in the model score is observed if a specific feature is critical to the predictive outcome. Little to no change in performance is recorded for non-influential features, when their values are randomized. The mean decrease in the score is calculated across fifty iterations to determine the final importance ranking. This approach is considered more robust than internal impurity measures because it is less biased by the scale of the variables.

Partial Dependence Plots

Partial Dependence Plots (PDPs) show how one predictor affects HILP predictions. The effect is calculated by averaging predictions across all other features. One feature varies across its full range while others remain fixed. For each value of the target feature, predictions are made across the dataset. All other features maintain their original values. The average prediction forms one point on the plot. This is repeated for every possible value of the target feature. The horizontal axis plots the target feature values. The vertical axis shows average predicted HILP probability. Smooth curves connect the average points. Confidence intervals are computed from prediction variability. The calculation of a PDP assumes that the feature of interest is independent of all other features in the model. This assumption is rarely met in real-world disaster datasets where socio-economic variables are often intertwined. Additionally, interpretations should be handled with caution in areas where the feature distribution is sparse. As data density decreases, the PDP curve is supported by fewer observations that makes the results increasingly unreliable. This lack of data density in extreme ranges is a key limitation and a caveat for the interpretation of the results.

Chapter 3 Outcomes of Unsupervised Machine Learning Models

The objective of adopting unsupervised machine learning models was to explore whether the algorithms can separate HILP events from non-HILP events so that we can understand which factors are influential in the context of HILP events. As discussed in the previous section, we utilized two unsupervised machine learning algorithms: self-organizing maps and K-means clustering. Both algorithms used three variables in forming the clusters. They are precedence, number of deaths per million population, and economic losses per billion GDP. These variables were normalized through a min-max normalization before using them in the machine learning models. In this section, we explain the outcomes of the two algorithms.

3.1. Self-Organizing Maps

Self-organizing maps project high dimensional data onto a 2-D grid, where each cell (neuron) represents a prototype vector, and nearby cells correspond to similar data patterns. The grid size can be derived from [18]. It states that the total number of cells in the grid (M) should be based on equation 1, where N is the total number of data points in the dataset. Based on this equation, the total number of cells, i.e., M should be $5 \times \sqrt{6211} \approx 394$. Therefore, for square grid, the grid size, i.e., the number of rows (or columns) should be $\sqrt{394} \approx 20$.

$$M = 5 \times \sqrt{N} \quad (1)$$

The outcomes of the SOM are shown in Figure 4.

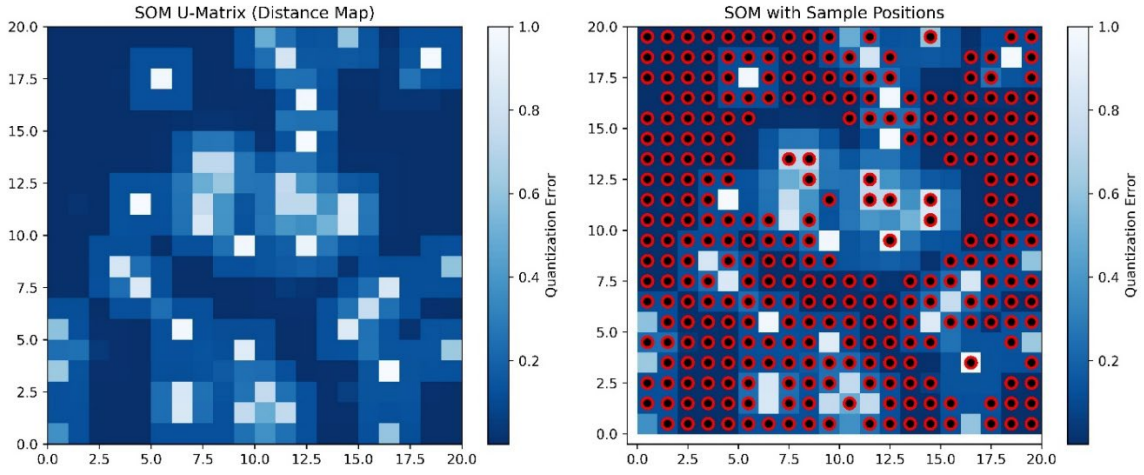


Figure 4 Self Organizing Map Output

The first plot on the left is presented as a Unified Distance Matrix (U-Matrix), where the topological structure of the data is visualized through the distances measured between neighboring neurons. Small distances between neurons are represented by dark blue regions, which indicate the presence of cohesive clusters where high similarity among data points is observed. Large distances are indicated by light blue and white areas, which serve as boundaries or separators between those identified clusters. Within the map, several distinct dark blue valleys are revealed, which are separated by a complex network of white ridges, suggesting that multiple subgroups are contained within the 6211 rows of data.

In the second plot, Sample Positions (represented by red and black circles) are overlaid onto the U-Matrix so that the distribution of actual data points across the grid can be observed. It is seen that the majority of the samples are concentrated within the dark blue, low-distance regions, which confirms that these areas function as high-density cluster centers. It is also noted that many of the white boundary cells are either left empty or are occupied by very few samples, as these regions are expected to represent the transitions between groups. However, a few isolated samples are found sitting directly on high-error white peaks. These specific points are likely categorized as outliers or unique anomalies that are not well-fitted into the primary clusters identified by SOM.

It must be noted that the exact number of existing clusters is not explicitly provided by the U-Matrix or the sample position plot. While dark blue regions are interpreted as potential clusters and white ridges are seen as boundaries, a definitive count of groups is not automatically generated by the SOM algorithm itself. Instead, the visualization is used as a tool for qualitative assessment, where the number of clusters is inferred by the observer based on the density and separation of the blue valleys. This means, final grouping would still depend on the visual interpretation rather than the concrete automated cluster results. Furthermore, the boundaries in the center of the map are presented as a complex and fragmented network, multiple interpretations regarding the total number of subgroups could be formed. To obtain a concrete numerical count, a secondary clustering technique, K-Means Clustering, was adopted. The use of K-means allows us to obtain reproducible partitioning of the dataset, which provides transparency and replicability. Therefore, SOM was retained as an exploratory tool to understand relations and topological structures, while K-means clustering was applied to derive definitive classification of HILP and non-HILP events.

3.2. K-means Clustering

To identify the number of clusters, i.e. K in K-means clustering, first, the elbow method was used. Figure 5 shows the outcome of the elbow method. The left panel in Figure 5

displays the elbow method. It shows the inertia, i.e., sum of squared distances from points to their cluster centroids against k values from 1 to 10. The curve starts high at k=1 (all points in one cluster) and drops sharply through k=2, reflecting rapid improvements in cluster tightness as groups separate. Beyond k=2, the decline flattens noticeably thus, forming the characteristic elbow around k=2. This pattern indicates that the dataset contains 2 clusters. Beyond that further subdivision captures noise rather than meaningful structure.

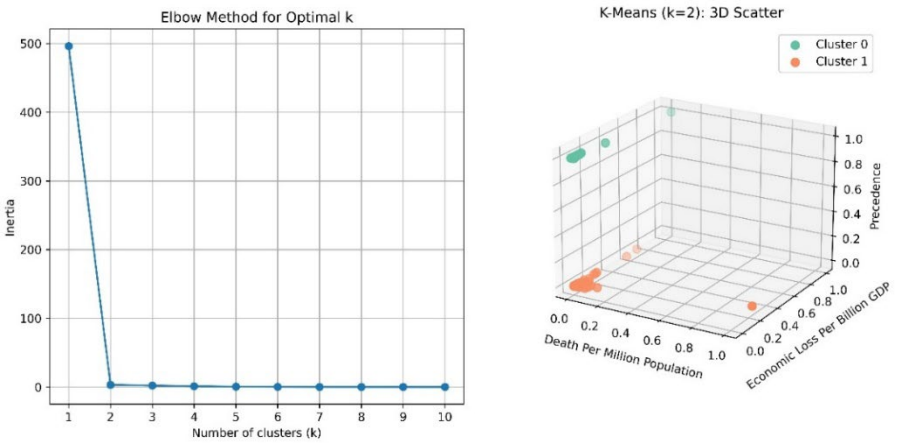


Figure 5 K-Means Clustering Elbow Method Outcome and Scatter Plot

The right panel presents a 3D scatter plot of the k=2 clustering solution. It visualizes how data points distribute across Death Per Million Population (x-axis), Economic Loss Per Billion GDP (y-axis), and Precedence (z-axis). The colors distinguish between the two clusters. The cluster concentrated on top (cluster 0) represents events with a precedence. In contrast, the disasters belonging to the bottom cluster (cluster 1) are without a precedence. The difference of impact parameters among the two clusters is not apparent in the scatter plot.

HILP and Non-HILP Separation

To make the difference more apparent, we compared the average death, damage, and precedence parameters between the two clusters in Figure 6.

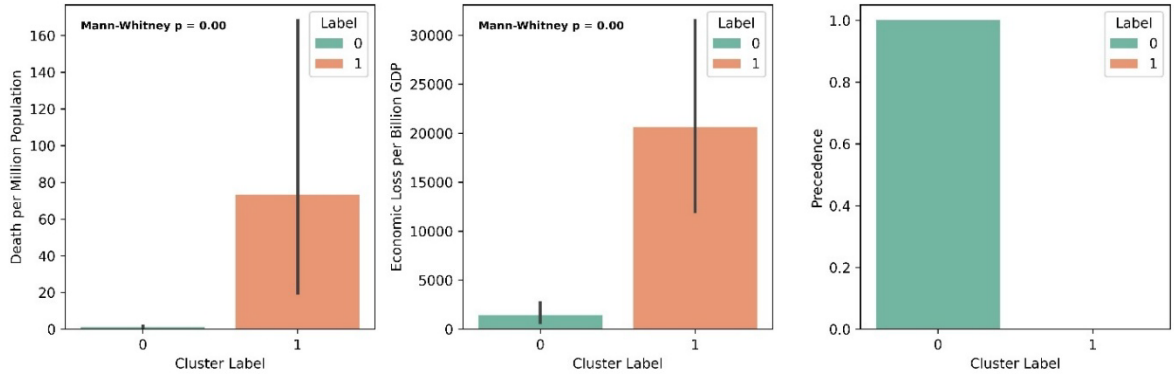


Figure 6 Average Death, Damage, and Precedence Difference between Two Clusters

It is apparent that the disasters belonging to cluster 1 experienced significantly higher deaths and economic losses. The average deaths per million population in cluster 1 was approximately 73 with a standard deviation of 995. On the other hand, the average deaths per million population in cluster 0 was approximately 1 with a standard deviation of 17. Similar differences were also noticed for the other impact variable, i.e., economic loss per billion GDP. The average economic loss per billion GDP among the disasters in cluster 1 was approximately $21000 with a standard deviation of $119000 whereas the same for the disasters in cluster 0 was $1433 with a standard deviation of $40000. The statistical significance of the inter-cluster difference of death and economic loss variables were also tested using the Mann-Whitney U test, which resulted in p-values less than 0.05 indicating that the differences are statistically significant.

Also, the disasters in cluster 1 lacked country-level precedence. This cluster represents 540 disasters out of 6211 in our dataset, i.e., 8.7% of the total events. On the other hand, the disasters belonging to cluster 0 that experienced much lower levels of deaths and damage had a country-level precedence. This cluster contains the remaining 91.3% of the data points. Based on this insight, we labelled the disasters belonging to cluster 1 as HILPs as they had significantly higher levels of impact and lacked precedence (our representative variable for probability). Similarly, the disasters belonging to cluster 0 were labelled as non-HILPs.

Characteristics of HILP Events

Figure 7 summarizes the distribution and characteristics of HILP disasters across hazard types, income groups, and leading countries by impact.

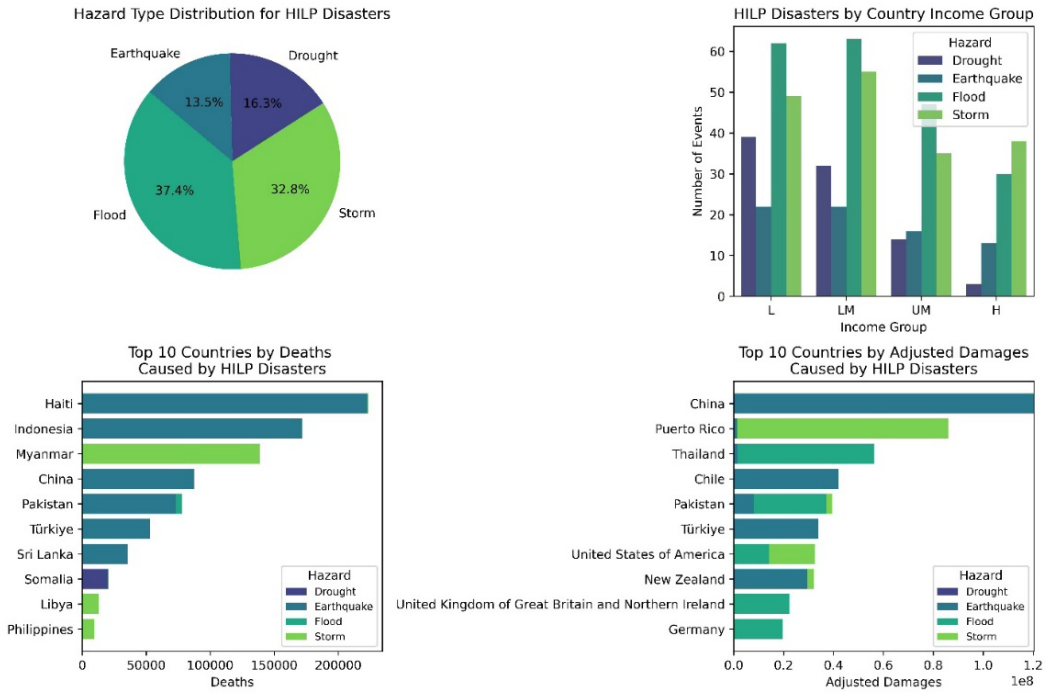


Figure 7 Characteristics of HILP-Labelled Disasters from K-means Clustering

On top left, the pie chart shows the hazard distribution among the HILP-labelled disasters. It can be noticed that approximately 70% of the HILP events are caused by floods and storms. Out of the remaining 30%, 16% are caused by droughts and 14% are caused by earthquakes. The top right plot shows the distribution of HILP events by country income groups. Among the low (L), lower-middle (LM), upper-middle (UM) income countries, most HILP disasters are caused by floods. On the contrary, among the high-income countries, most HILP disasters are caused by storms. HILPs caused by droughts decrease as countries grow richer. A similar trend can be observed for earthquakes as well.

The geographical variation of these events is visualized through stacked horizontal bar charts for deaths and adjusted damages. Each country's total is segmented by hazard type, allowing for a clear comparison of the primary drivers of loss in each region. The inverted y-axis ensures that the most severely impacted nations are positioned at the top for immediate identification. Haiti and Indonesia are identified as the countries with the highest mortality rates, largely driven by earthquakes. It can be noticed that 7 out of the top 10 countries that suffered the most fatality due to HILP disasters are from Asia. In terms of economic impact, China is shown to have the highest adjusted damages, predominantly caused by earthquakes. Other nations such as Puerto Rico and Thailand show significant losses driven by storms and floods, respectively. It is also interesting to see the 3 largest economies in the world, i.e., the USA, China, and Germany suffered a lot of economic losses from HILP disasters.

Figure 8 further shows how a wide range of socioeconomic and infrastructural indicators differ between non-HILP (label 0) and HILP (label 1) disasters.

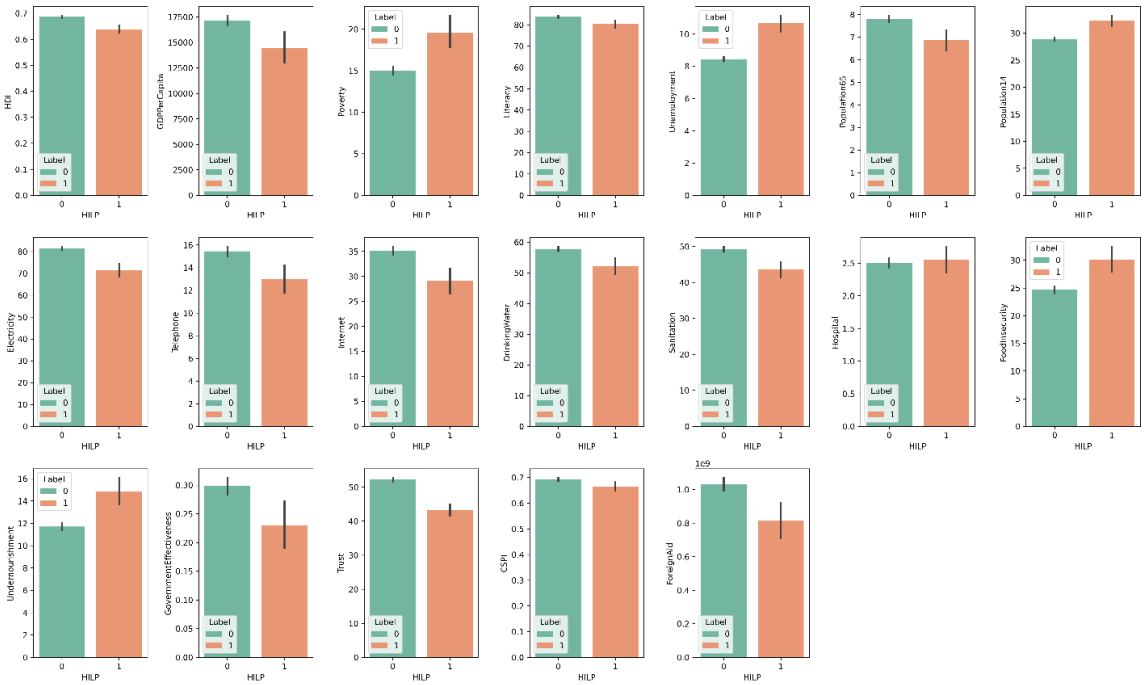


Figure 8 Differences of the Dimensions of Disaster Risk between HILP and Non-HILP Disasters

Each panel displays the mean value of one predictor by label, with non-HILP bars in green and HILP bars in orange and with error bars indicating uncertainty around the mean. By reading across the panels, a profile of the typical context in which HILP disasters occur can be inferred. Differences in level and direction between the two bars in each plot indicate whether HILP events are associated with more or less favorable country conditions for that variable.

In the first row, lower HDI values are observed for HILP events, while GDP per capita also appears lower for HILP compared with non-HILP contexts. Poverty levels look slightly higher or similar under HILP, though the difference is modest. Literacy is shown as marginally lower for HILP countries, suggesting weaker human capital. Unemployment levels appear somewhat higher for HILP labels. Population older than 65 seems slightly higher for HILP events, indicating a more vulnerable demographic composition for HILP contexts.

In the second row, access to electricity is lower for HILP than for non-HILP disasters. Telephone and internet penetration are also clearly lower for HILP, representing weaker communication infrastructure. Access to drinking water and sanitation is reduced for HILP events, which implies poorer basic services and higher baseline vulnerability. Hospital availability per capita appears slightly higher under HILP, but the difference is small and may reflect measurement noise or targeted health investment in risk prone settings. Food insecurity is higher for HILP, consistent with more fragile livelihoods and weaker safety nets.

In the third row, undernourishment is higher in HILP contexts, reinforcing the picture of chronic vulnerability. Government effectiveness is lower for HILP labels, pointing to weaker state capacity to manage risks and respond. Trust in institutions is likewise lower, which may hinder preparedness and compliance with risk reduction measures. The composite CSPI indicator is slightly lower for HILP, suggesting an overall less favorable social environment. Foreign aid appears lower in HILP countries, indicating that these states are more aid dependent and may lack domestic resources for resilience and disaster risk reduction measures.

Taken together, HILP disasters are characterized by occurrence in countries with lower income, lower human development, and weaker service provision. Structural vulnerabilities are reflected in poorer access to infrastructure, water and sanitation, and higher food insecurity and undernourishment. Governance-related indicators show reduced government effectiveness and trust, which implies limited institutional capacity to anticipate and manage extreme but rare events. HILP disasters are therefore seen as concentrated in more fragile, lower-capacity settings, where exposure and vulnerability are jointly elevated, while non-HILP disasters occur more often in better resourced and better governed environments.

Chapter 4 Outcomes of Supervised Machine Learning Model

The unsupervised machine learning model found 2 clusters of disasters from the dataset. The first cluster with 8.7% of the data points lacked precedence but experienced significantly higher fatality and damage than the other cluster, which had precedence. Therefore, the first cluster was considered as HILP disasters, and the remaining 91.3% disasters were considered as non-HILP disasters. After we established this segregation of HILP and non-HILP disasters, we utilized a supervised machine learning model through Random Forest (RF) classification algorithm to explore to what extend these labels can be predicted by variables relevant to the dimensions of disaster risk.

The variables have already been explained in section 2.4. The missing values among the predictors were imputed using k-nearest neighbor algorithm. Before using the predictors in the RF model, all variables were normalized through min-max normalization to eliminate the effects of differences in variables' ranges. We also had to perform synthetic oversampling of the minority class due to severe class imbalance in the dataset. The details of minority oversampling can be found in Appendix B.

To summarize, the predictors and the response variable were pre-processed differently before training the RF classification model. For predictors, the missing values were first imputed using k-nearest neighbor algorithm. Next, the variables were normalized with min-max normalization. After that, the data was split into training (80%) and test set (20%). Finally, on the training set, the Synthetic Minority Over-sampling Technique (SMOTE) algorithm was implemented to produce a balanced synthetic training set that added synthetic data to the original training set. The RF model was trained on this balanced synthetic training set. The hyperparameters were tuned using a grid search method, which has been explained before. The optimal RF model had 300 trees with a maximum depth of 50, and a minimum of 20 samples required for splitting.

4.1. Classification Results

The performance of the classification model was assessed on the held-out test set, which contained 20% of the dataset, i.e., 1243 datapoints. The outcomes of the classification model are represented as a confusion matrix, which is shown in Figure 9. The confusion matrix summarizes classification performance by comparing true HILP labels against model predictions on the test set. True non-HILP events (label 0) occupy the top row, while true HILP events (label 1) occupy the bottom row. Correct predictions appear on the

diagonal: 1071 non-HILP and 70 HILP events were identified accurately. Off-diagonal cells show prediction errors: 64 non-HILP events were falsely labeled HILP (false positive), and 38 HILP events were missed as non-HILP (false negative).

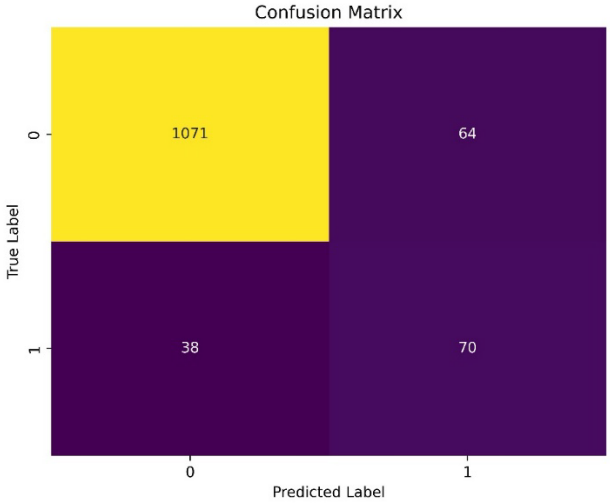


Figure 9 Confusion Matrix

The performance of the classification model is expressed through 4 class-wise metrics shown in Table 2. These 4 metrics are derived from the confusion matrix. Non-HILP events achieved precision of 0.97, recall of 0.94, and F1-score of 0.95. HILP events showed precision of 0.52, recall of 0.65, and F1-score of 0.58. Overall accuracy reached 0.92, with 1141 correct predictions out of 1243. Macro-averaged metrics were also computed as precision 0.74, recall 0.80, and F1-score 0.77. Weighted averages aligned closely with accuracy at 0.93, 0.92, and 0.92 respectively. These results indicate excellent discrimination of typical non-HILP disasters but moderate performance in identifying rare HILP events, consistent with class imbalance challenges.

Table 2 Classification Performance

Label	Precision	Recall	F1-score	Datapoints
0 (Non-HILP)	0.97	0.94	0.95	1135
1 (HILP)	0.52	0.65	0.58	108
Accuracy			**0.92**	1243
Macro average	0.74	0.80	0.77	1243
Weighted average	0.93	0.92	0.92	1243

These results imply that typical, non-HILP disasters are characterized and separated well by the chosen predictors and modeling pipeline, whereas HILP disasters are harder to distinguish. The moderate recall for HILP indicates that a nontrivial fraction of high impact low precedence events would still not be flagged by the model, which may be problematic

for early warning or prioritization purposes. At the same time, the limited precision suggests that many events flagged as HILP would in fact be non-HILP.

4.2. Variable Importance Analysis

Feature importance scores from the Random Forest classifier are displayed as horizontal bars, ranked by their contribution to distinguishing HILP from non-HILP disasters in Figure 10 and 11. Figure 10 shows the impurity-based feature importance whereas Figure 11 shows the permutation-based feature importance.

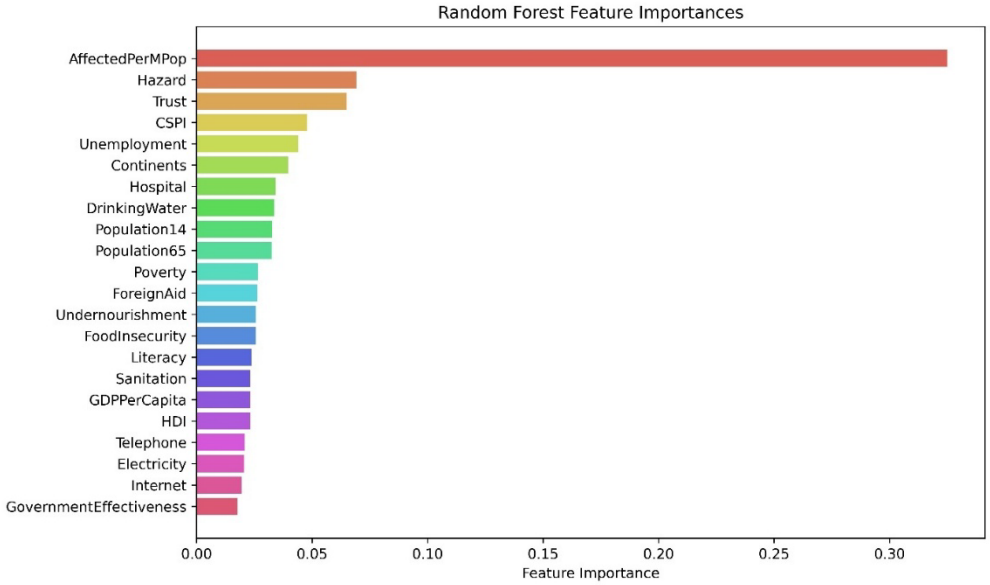


Figure 10 Variable Importance Plot (Impurity Based)

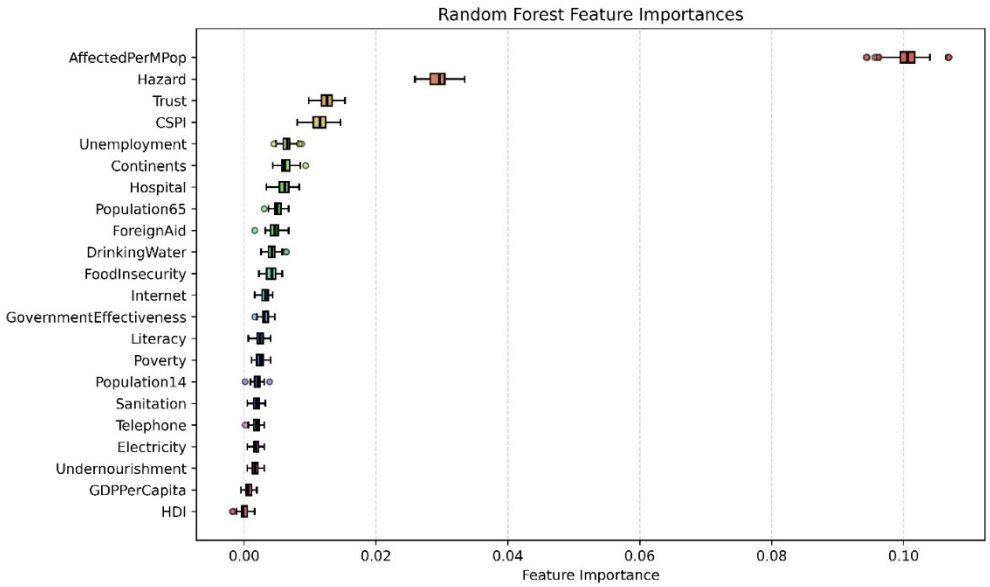


Figure 11 Variable Importance Plot (Permutation Based)

The order of importance between the two plots is very similar proving the robustness of outcomes. In both plots, AffectedPerMPop, which measures the *number of people affected*

by the disaster per million population of the country, i.e., the representative variable for exposure, is shown as the dominant predictor at the top. It is followed by *Hazard*. Thus, *Hazard* and *Exposure* play the most significant role in distinguishing HILP disasters from non-HILPs.

Trust, Unemployment, and *CSPI* follow closely. The relatively high importance of *trust* suggests that the level of public trust in government, institutions, etc., plays a role in disaster outcomes. It affects how well people follow evacuation orders or how efficiently aid is distributed. Elevated *unemployment* signals weaker household resilience and limited coping capacity. High importance of *CSPI* indicates that community engagement is also an important determinant of catastrophic disasters.

Macroeconomic metrics, including *GDP per capita* and the *Human Development Index*, are found to have surprisingly limited predictive power. This result is interpreted through another research's finding that HILPs are not restricted to any specific economic development level [1]. The lowest ranked variables are shown to be infrastructure metrics such as electricity and internet access. These factors are considered less discriminatory for identifying HILPs compared to the immediate human scale of the event.

Government effectiveness is also positioned at the lower end of the importance hierarchy. This implies that HILPs, by their nature as outliers, are expected to overwhelm even robust governance mechanisms. The model is thus driven by the interaction between extreme triggers and the underlying social fabric. The classification of HILPs is shown to rely on the nuanced relationship between a hazard and the specific vulnerability of the affected population.

4.3. Partial Dependence Plots

To get a better understanding of how these variables influence the HILP and non-HILP classification, we produced the partial dependence plot. Before creating the plots, we tested the pairwise correlations among the predictors. The correlation matrix is shown in Figure 12. High pairwise correlations can be noticed among the socio-economic variables such as *per capita GDP, HDI, poverty, literacy, population over 65, population under 14*, access to infrastructures *electricity, telephone, internet, drinking water, sanitation, and hospitals, food insecurity, and undernourishment*. Interestingly, the predictors that ranked highest in variable importance analysis, i.e., *affected people per million population*, *trust, CSPI, and unemployment* did not have high pairwise correlations (coefficient higher than 0.50) with other predictors.

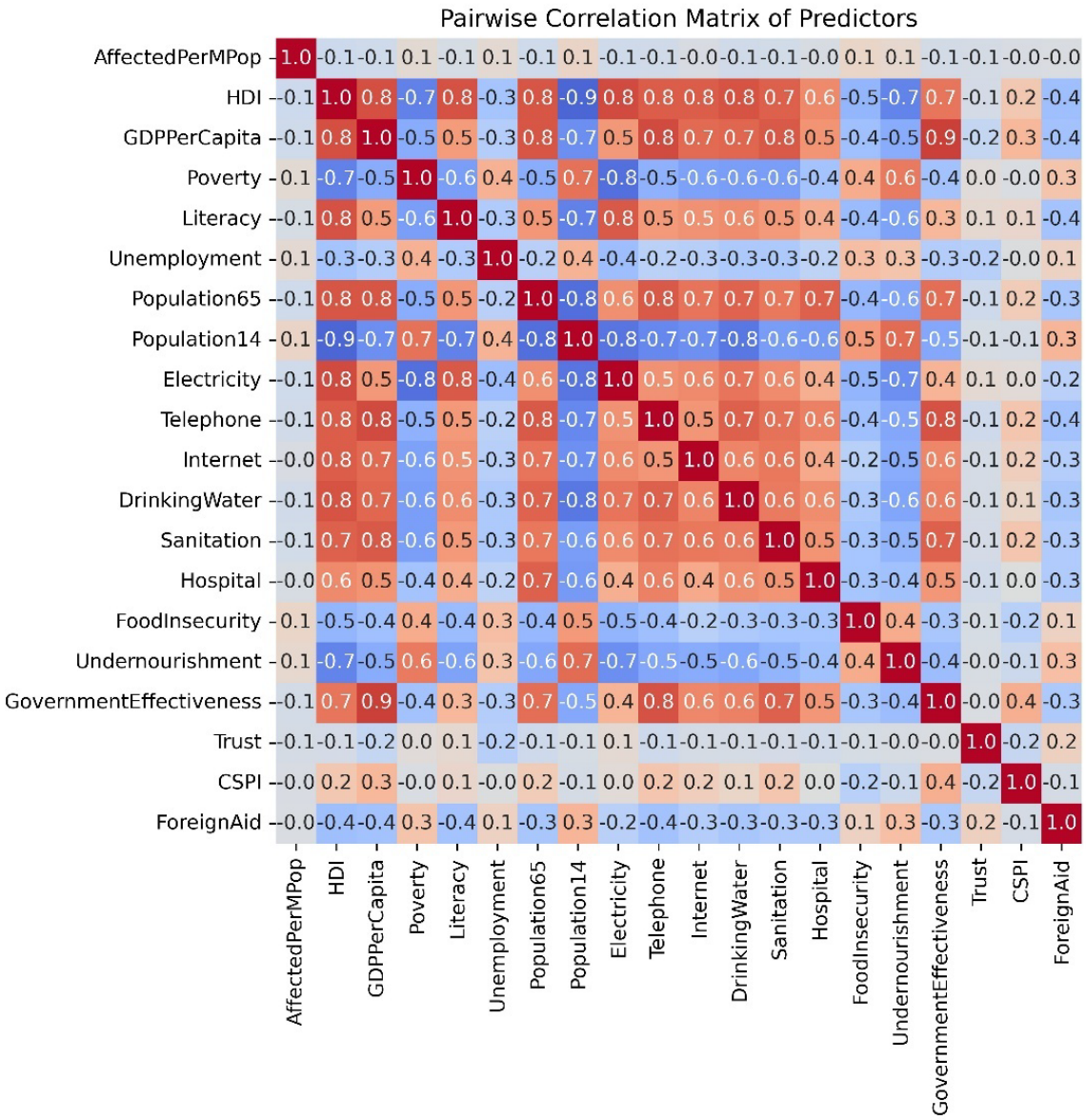


Figure 12 Pairwise Correlations among Predictors

These partial dependence plots are shown in Figure 13. The plots are created for 20 predictors that exclude two categorical variables, i.e., *hazard* and *continent.* Due to low pairwise correlation, the PDPs of *affected people per million population*, *trust, CSPI, unemployment, and foreign aids* are more accurate than the others.

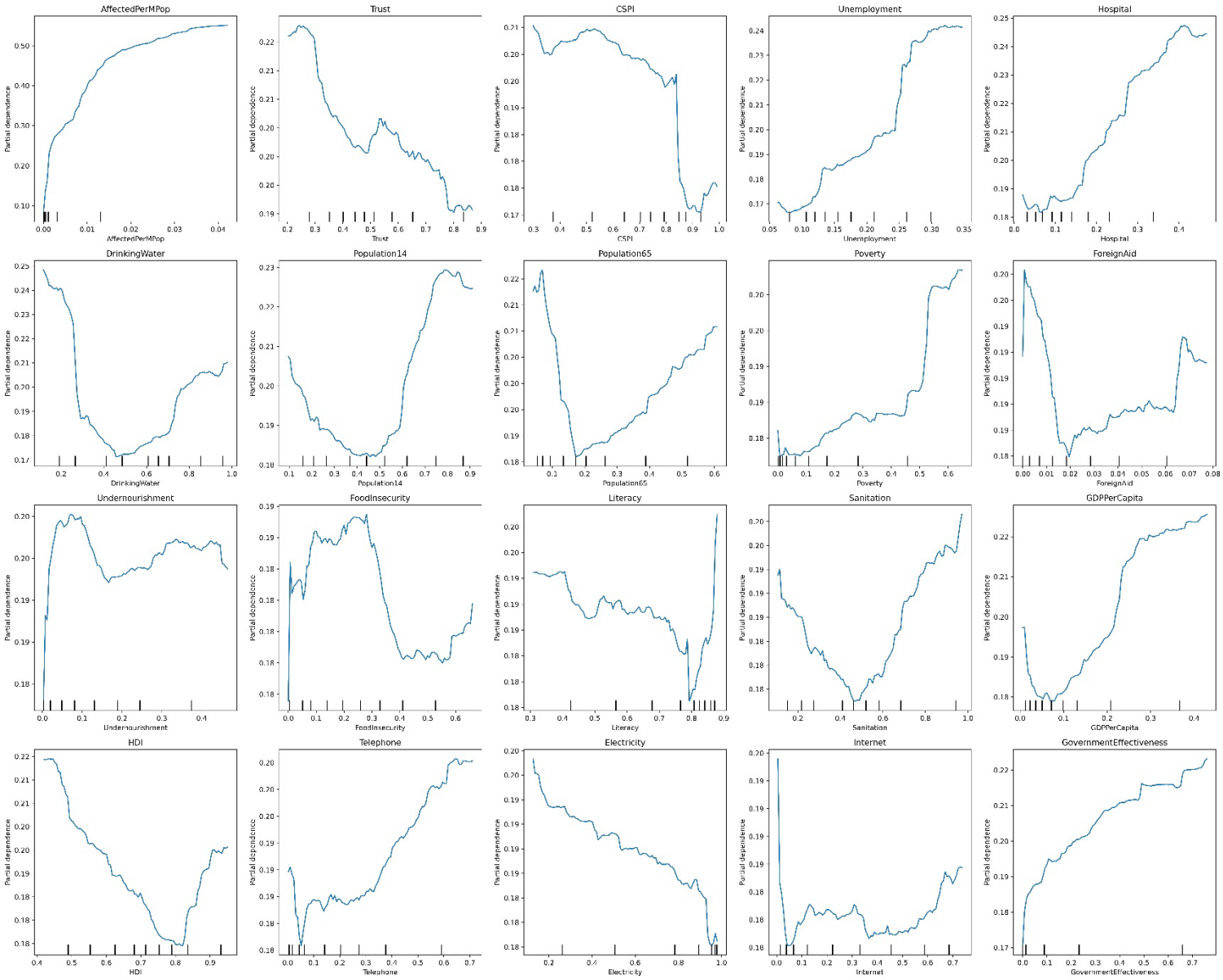


Figure 13 Partial Dependence Plots

The relationship between ***the proportion of the population affected*** and the probability of HILP classification is characterized by a sharp, early increase. This steep gradient suggests a critical threshold beyond which a disaster is almost certainly classified as a HILP event. Such a pattern is consistent with past classification that considered HILPs as events that produced catastrophic consequences relative to the affected population [1]. The plateau observed at higher values indicates that once a specific level of systemic impact is achieved, further increases in the affected population do not significantly change the classification probability. This reflects the outlier nature of HILPs where the scale of impact is inherently extreme.

A significant inverse relationship is observed between ***social trust*** and the likelihood of a HILP event. Higher levels of trust are associated with a substantial decrease in the probability of a disaster escalating into a HILP outcome. Lower trust environments may suffer from failures in regulation and ignored warnings, which are regarded as key drivers for cascading failures in complex systems. High trust is interpreted as a marker for societal resilience and effective risk communication, both of which serve to mitigate the propagation of systemic shocks.

The partial dependence plot for ***GDP per capita*** reveals an unexpected positive correlation with HILP probability at higher income levels. This phenomenon is explained by the concept of normal accidents in tightly coupled, complex systems discussed in [1]. Highly developed economies often rely on interconnected infrastructure networks that are inherently prone to unanticipated interactions and failures. As systems become more complex and interdependent, they become more susceptible to cascading dynamics that can trigger HILP events regardless of economic wealth. This decoupling of economic development from disaster resilience highlights the vulnerability of modern, high-tech societies to systemic outliers.

Socio-economic vulnerability is further explored through the variables of ***unemployment*** and ***poverty***. An increase in the unemployment rate is shown to correlate with a higher probability of HILP classification, with a notable spike appearing as rates exceed twenty percent. High unemployment and poverty are viewed as indicators of low socio-economic resilience, which allows localized disruptions to escalate into systemic disasters. The demographic plots for ***younger and older populations*** also show increased risk at extreme percentages, reflecting the specific vulnerabilities of dependent groups in the face of cascading failures.

Infrastructure and service access plots, such as those for drinking water and electricity, demonstrate the protective nature of robust public services. A decrease in access to these critical services is linked to a higher probability of HILP outcomes. In countries where infrastructure is already fragile, the threshold for a trigger to cause a systemic collapse is significantly lowered. The non-linear drop in probability as service access improves

suggests that a baseline level of infrastructure is required to prevent disasters from reaching outlier status.

Government effectiveness is shown to have a monotonic relationship where higher effectiveness scores are generally associated with higher HILP probabilities. The probability remains non-zero even at high levels of effectiveness, which is consistent with past findings that HILPs can overwhelm even robust governance structures [1]. Outlier events, by definition, fall outside the standard scenarios for which most governments are prepared.

In summary, the partial dependence plots provide a visual representation of the complex, systemic drivers of HILP events. The prominence of threshold effects and the influence of social factors like trust and economic complexity mirror the theoretical frameworks presented in literature. The model correctly identifies that HILPs are not merely the result of strong physical triggers but are emergent properties of vulnerable, tightly coupled systems. These plots emphasize that mitigating HILP risks requires addressing the underlying societal and structural criticalities that allow impacts to cascade into catastrophic outliers.

4.4. Guidelines of HILP Events and Recommended Mitigations

Based on the outcomes of the supervised machine learning model and the analysis of HILP events, the following mitigation measures are proposed:

- **Population Exposure Monitoring and Reduction**: Affected population per million was found as a primary indicator for identifying potential HILP events. Higher population exposure was found to monotonically increase the probability of turning a disaster into a HILP event. This non-linear spike is a consistent trend across all hazard types studied as shown in Figure 14.

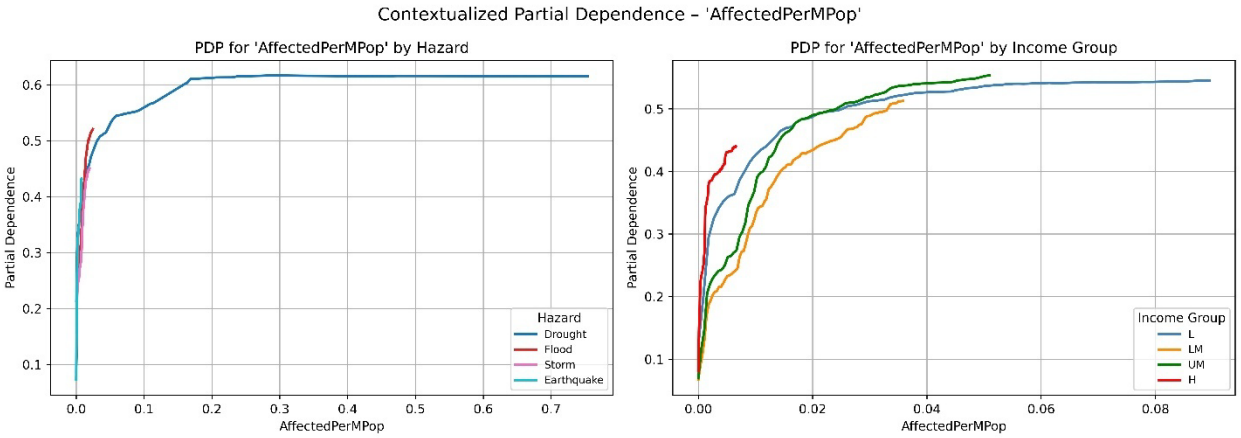


Figure 14 Hazard and Context Specific PDP of Population Exposure

Earthquakes and storms show the most immediate risk of escalation at very low exposure levels. Droughts exhibit a more gradual increase but maintain a high-risk plateau once

exposure is high. High-income countries (H) show the highest sensitivity to population exposure among all income groups: high (H), upper middle (UM), lower middle (LM) and low (L). These nations reach maximum HILP probability at significantly lower exposure levels than their lower-income counterparts. Due to its significance, **real-time exposure tracking** should be adopted and implemented to trigger emergency responses at the earliest signs of mass impact. Additionally, mitigation measures such as **building dams and barriers** (reduces exposure from floods and storms), **adoption of earthquake resistant building codes** (reduces exposure from earthquakes), etc. should be prioritized to reduce population exposure to natural hazards like floods, storms, and earthquakes.

- **Trust-Based Communication Protocols**: Institutional trust strengthens the efficacy of risk communication during a crisis. Evacuation compliance is facilitated through increased public confidence in government and relief institutions. Hence, institutional trust becomes a powerful driver of resilience especially in low and lower middle-income nations as shown in Figure 15.

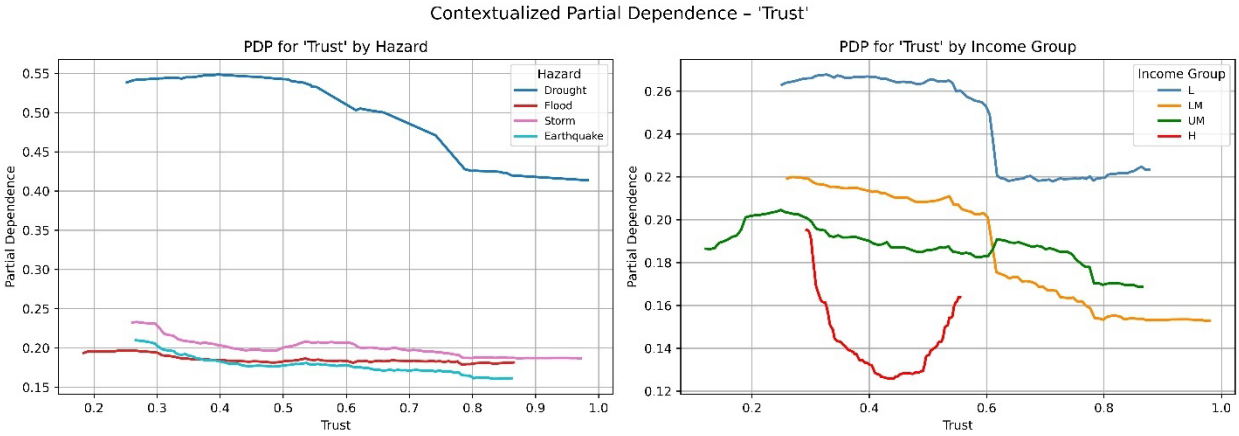


Figure 15 Hazard and Context Specific PDP of Institutional Trust

A sharp decline in HILP probability occurs when trust levels surpass the 0.6 threshold in these contexts. High-income societies show a more complex relationship with an optimal risk reduction point at moderate trust levels. Both low and high levels of trust are associated with increased HILP probability in high income societies. Drought is the hazard most affected by trust levels with a sharp decline in risk starting at a level of 0.5. Floods and storms show a much flatter relationship but still benefit from increased public confidence. Due to the importance of institutional trust in mitigating the probability of a disaster turning into HILP event, authorities and policy makers should strive to ensure that people trust their government and public institutions. This can be achieved through **transparent risk communication, accountable distribution of relief resources, and the active inclusion of local community leaders** in the emergency planning process. Such measures are essential to reaching the 0.6 trust threshold, which the model identifies as a critical tipping point for ensuring systemic stability during a crisis.

- **Targeted Socio-Economic Resilience**: Socio-economic factors such as poverty, unemployment, and others reduce household-level resilience as evident from Figure 16.

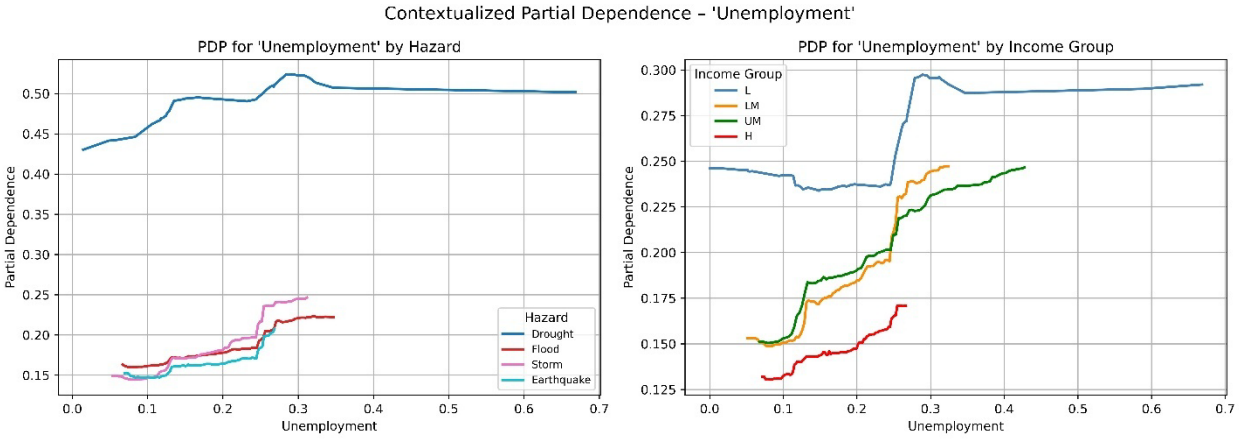


Figure 16 Hazard and Context Specific PDP of Unemployment

Limited coping capacities are mitigated through targeted social safety nets and financial assistance. Unemployment particularly acts as a critical threshold for disaster escalation with a major tipping point observed at unemployment level of 0.25. Drought-related risks are particularly sensitive to labor market instability and maintain a high baseline probability. Floods, storms, and earthquakes show a distinct stepwise increase in risk as unemployment rates move from **0.2** to **0.3**. Low-income nations face the highest baseline risk, but all income groups experience a sharp rise at similar unemployment levels (approximately at the level of 0.1). High-income countries exhibit a lower baseline, but a steep increase in vulnerability once unemployment exceeds 0.2. These patterns suggest that economic precarity drastically lowers a system's ability to absorb shocks **against HILP events**.

As mitigating measures, governments should establish automatic triggers for **social safety nets and emergency income support** that activate as regional unemployment approaches these identified tipping points. In drought-prone regions unemployment reaching the 0.25 threshold, mitigation should prioritize **labor market stabilization and unemployment insurance schemes** to ensure households do not lose their primary coping mechanisms during prolonged crises. For high-income contexts, proactive **employment protection and retraining programs** in sectors vulnerable to cascading failures can prevent localized economic stress from escalating into a systemic HILP event.

- **Community Engagement and Social Protection**: Community engagement indices like CSPI should be monitored to evaluate the strength of local resilience. CSPI can be used as a determinant to gauge the readiness of communities for catastrophic **events**. It has been found that higher CSPI reduces the probability of a disaster turning into a HILP event in all country income groups. Therefore, efforts must be made to

increase the participation of civil society into legislation. Figure 17 shows the hazard and context specific PDPs for CSPI.

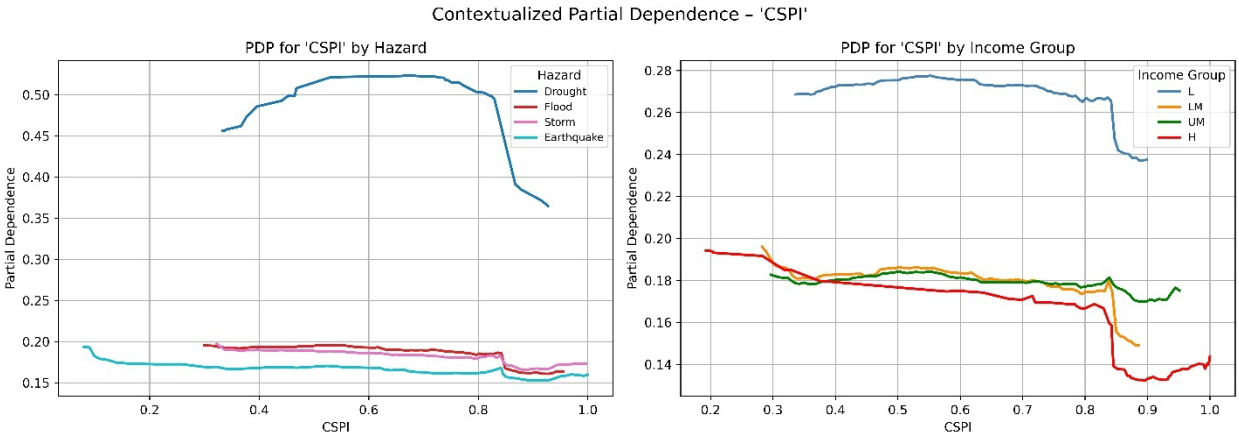


Figure 17 Hazard and Context Specific PDP of Community Engagement

For all hazard types and income groups, a CSPI above 0.8 serves as a critical tipping point. Below this threshold, communities remain highly susceptible to systemic escalation, particularly in economically vulnerable contexts. Policies should prioritize reaching this zone, i.e., CSPI higher than 0.80 to ensure that social infrastructures can absorb extreme shocks from HILP disasters.

As mitigating measure, authorities should implement **community-led disaster management programs** that grant residents a direct voice in local resilience strategies. Formally **integrating civil society organizations** into the legislative process for risk management ensures that grassroots social networks are empowered to act as a critical buffer against catastrophic escalation. This bottom-up approach transforms passive residents into active participants, creating the high-density social fabric required to absorb shocks that might overwhelm traditional top-down governance.

- **Stress-Testing of Complex Infrastructure**: Infrastructure networks particularly in high-income economies should be **stress-tested against cascading failures**. Due to their interconnectedness, even small events can turn into catastrophic disasters. Hence, **interdependent network risks should be monitored** to ensure resilience in highly connected societies. Our partial dependence results show that the characteristics of hazards need to be also taken into account in stress testing, as the effects of slow-onset events such as draughts might be different than sudden-onset events such as floods, storms and earthquakes.
- **Maintenance of Critical Service Baselines**: Minimum baseline access to essential services, such as electricity and drinking water, should be maintained as a primary protective measure. Robust public service provision must be ensured to prevent localized disruptions from escalating into systemic HILP events. **Higher access to infrastructure services** increases the threshold at which a disaster becomes a HILP

event. For specific hazard types such as sudden-onset events, maintenance should target reducing immediate impacts, for slow-onset events such as draughts maintenance should target long-term efficiency.

- **Threat-Agnostic Governance Adaptation**: Risk management frameworks need to be designed to be threat-agnostic to address outliers that exceed standard governance capabilities. Governance systems should **be designed to remain functional under extreme conditions that exceed conventional operational conditions**. The non-linear relationship between trust and HILP risk in high-income contexts (see Figure 15) indicates that rigid, centralized governance is often insufficient. Instead, policy makers should adopt **adaptive and flexible arrangements** that can pivot as social conditions change. These frameworks must be supplemented by **threat-agnostic strategies** to manage extreme scenarios that fall outside traditional risk registers. By building institutional agility, governance systems can withstand triggers that overwhelm standard bureaucratic protocols.
- **Demographic-Specific Response Protocols**: **Demographic risk assessments** and specific response strategies need to be integrated in disaster planning. Specifically, vulnerability assessments for the young and the elderly need to be conducted to reduce the HILP probabilities. **Targeted response protocols** should be developed for regions with very high percentages of dependent populations (e.g., high elderly and dependent populations). Susceptibility to cascading failures can be reduced through targeted protection of the most vulnerable age groups. For sudden-onset disasters, targeted communication strategies might be essential while for slow-onset events such as draughts long term support of the communities are more critical.

Chapter 5 Validation

In task T2.3, we implemented a two-stage machine learning framework to uncover the drivers of the HILP events. The unsupervised machine learning algorithm, i.e., K-means clustering labelled approximately 9% of the historical disasters as HILP events. Next, we utilized the Random Forest classification model that found that *population exposure, hazard, institutional trust, community engagement*, socio-economic factors such as *unemployment* are the five most influential factors that explain why some events turn out to be HILP disasters. It is important to note that the outcomes are based on data-driven analysis. Therefore, they require some form of validation. We used several validations to support the outcomes.

5.1. Validation via Stakeholder Survey (Task T2.1)

Within AGILE project's task T2.1, we conducted a stakeholder survey between May 2, 2024 and June 18, 2024, where we asked experienced practitioners who have managed disasters caused by several hazard types around the globe to identify the key determinants of HILP disasters. We identified ten such determinants from literature and through the survey, we asked them to identify the top five determinants. The outcomes of the survey have previously been shared in deliverable D2.1. Figure 18 summarizes the input from 104 survey respondents. In deliverable D2.1, the survey design, data collection, and analysis have been explained thoroughly.

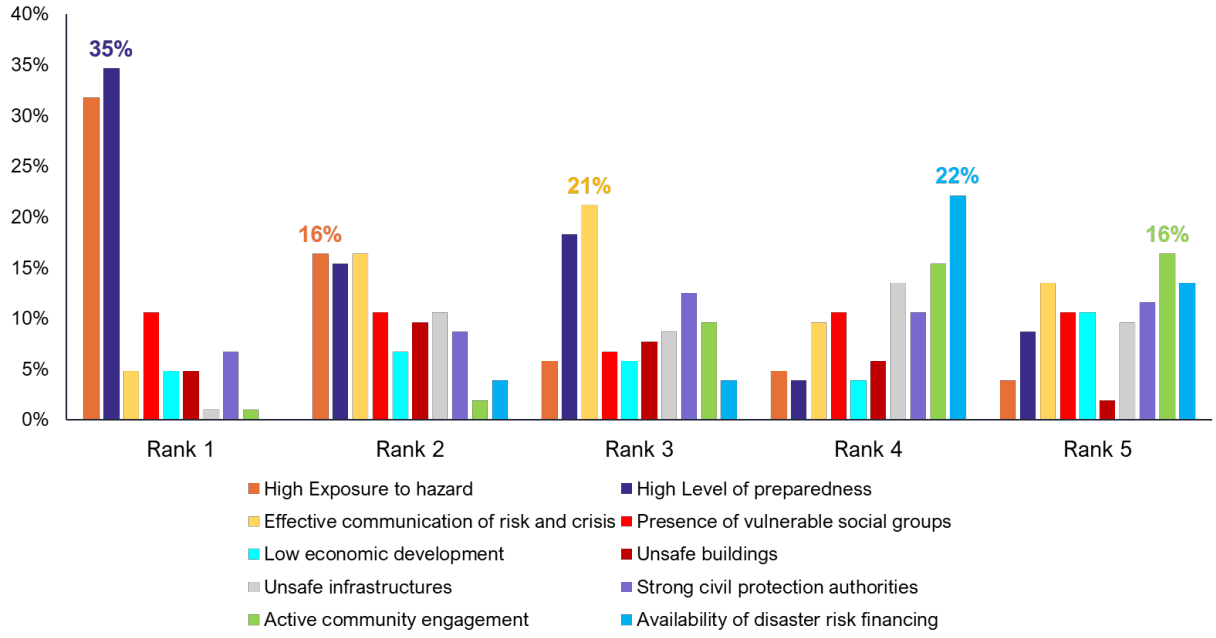


Figure 18 Outcome of Stakeholder Survey from T2.1

It can be noticed that surveyed stakeholders also identified *exposure to hazards* and *community engagement* as two of the five priority factors. From the Random Forest model, we found that *people affected per million population* (our representative variable for population exposure), *hazard*, and *CSPI* (our representative variable for community engagement) are among the top five predictors. Therefore, among the top five factors that influence HILP outcomes identified in task T2.1 and T2.3, we found 3 common factors. This significant convergence serves as a robust validation of the machine learning framework. It demonstrates that the algorithm is capturing genuine, real-world dynamics rather than just statistical artifacts. The agreement between historical data patterns and expert consensus suggests that these shared factors represent the fundamental drivers of HILP events that are independent of the analytical methodology applied.

5.2. Methodological Validation via Stakeholder Consultation

Unlike the outcome-based validation discussed in the previous section, this phase focused on validating the structure of the two-stage machine learning framework itself. We subjected the methodology to review on two specific occasions.

First, the framework was presented to internal project partners at the 4th AGILE General Assembly in Rotterdam (September 2025). This session detailed the data pipeline, including collection, imputation, cleaning, and the specific machine learning algorithms employed. Feedback from partners highlighted the need to analyze misclassifications in the Random Forest algorithm to detect systematic errors. Subsequent analysis revealed a geographic bias, where most misclassified events originated from a single continent. To address this, we introduced 'Continent' as a categorical variable. This allowed the model to account for region-specific latent features that were previously missed. Second, the framework was presented to a group of external experts at the International Climate Resilience Conference in Munich (October 2025). The methodology received largely positive feedback from the session attendees.

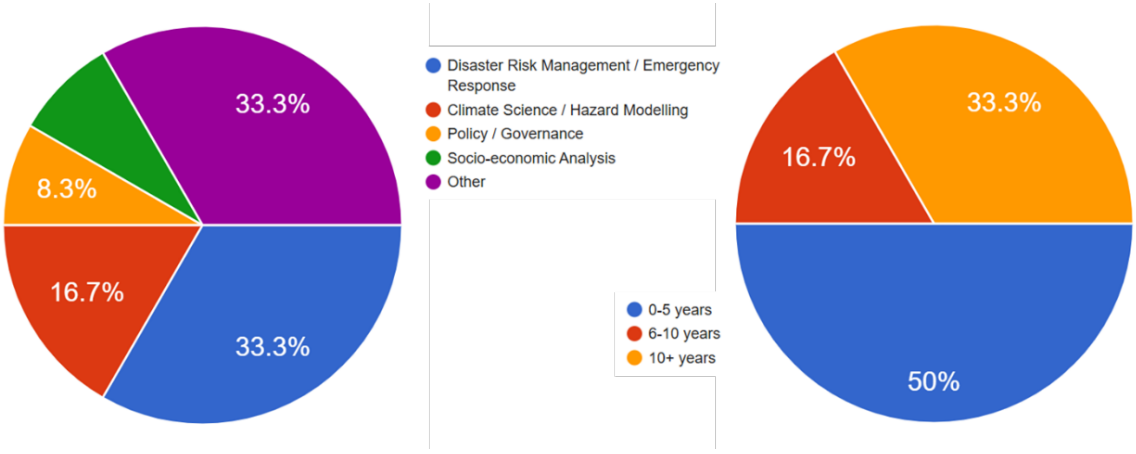
5.3. Validation Survey

The primary purpose of the stakeholder validation survey was to perform a cross-check on the outputs of the two-stage machine learning framework developed in Task 2.3. While the K-means clustering and Random Forest models provided objective, data-driven insights into the drivers of HILP events, it was essential to determine if these statistical correlations resonated with the lived experience of disaster risk practitioners. By seeking feedback from experts within the AGILE network, we aimed to bridge the gap between computational modeling and operational reality, ensuring that the variables identified by

the algorithm such as institutional trust, population exposure, etc., are viewed as actionable and relevant by those managing disasters on the ground.

The survey was designed to be concise and completely anonymous to encourage candid feedback from a diverse group of specialists within the project's network. The questionnaire was developed through Google Forms and can be found in Appendix C. The questionnaire was structured into four distinct thematic sections. The first section established the professional profile and seniority of the respondents. The second section required experts to rate and rank the HILP drivers identified by the Random Forest model shown in Figure 10 and 11. The third section prompted a methodological critique of the data-driven approach versus traditional expert judgment. The final section was dedicated to any open feedback from the respondents.

The data collection started on 24 February 2026 during an AGILE consortium meeting. First, the objectives of T2.3, methodology, and outcomes were presented through a brief presentation. This was followed by sharing the link to the online survey with the meeting participants. The survey was open until 1 March 2026. Twelve experts from various sectors with different levels of experience participated in the targeted survey as shown in Figure 19.



(a) (b)

Figure 19 Respondents' Backgrounds (a) Profiles (b) Professional Experiences in Years

The pool included specialists in disaster management, climate science, and policy. One-third of the respondents focus specifically on emergency response. Senior experts with over five years of experience represented 50% of the group. This balanced distribution captures both intuitive field judgment and data-driven perspectives. Overall, the diverse group ensures a robust validation of the machine learning framework across multiple professional silos.

Figure 20 represents the levels of agreement and/or disagreement among the survey respondents regarding the findings of the ML models. We asked the respondents about the top five factors we identified from the Random Forest model (see Figure 10 and 11).

The objective was to understand how much they agreed with these 5 factors. From Figure 20, it can be noticed that the respondents strongly agreed that population exposure is an important predictor of HILP events. For hazard type, we can notice a tie between agree and strongly agree. For the other three factors, we can see from Figure 16 that the respondents agreed with the supervised machine learning model's outcomes.

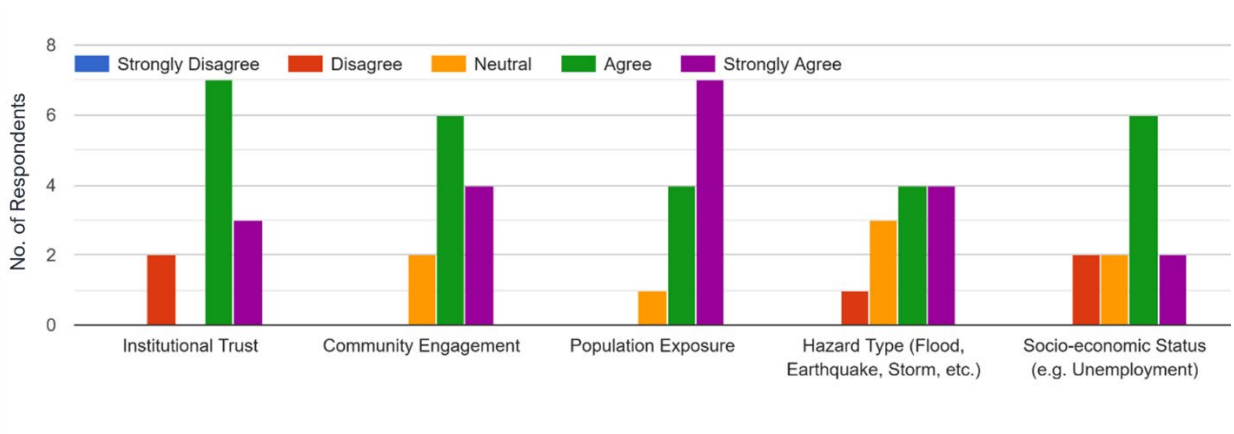


Figure 20 Agreements and Disagreements with Outcomes of Machine Learning Models

We also asked the respondents if they could pick only one of the five factors shown in Figure 20, which one they would pick. This was done to understand their preferences among the top five predictors from the Random Forest model. Nearly 42% selected population exposure as the most important predictor among the five, which was followed by Hazard type. From Figure 10 and 11, we can notice that these factors also came up at the top in terms of their importance to the Random Forest model's outcomes, thereby confirming further validation of T2.3 outcomes.

Lastly, we asked the respondents if they would prefer data-driven approaches over expert judgement for identifying drivers of HILP events like we did for T2.3. Two thirds of the respondents agreed that the data-driven approach in T2.3 is more useful whereas the remaining one third of the respondents claimed that expert judgements are more practical. This provided validation for task T2.3's methodology as well.

The open feedback emphasized the critical role of risk governance and management in preventing HILP events. Experts highlighted the importance of accounting for compound and cascading crises, particularly in conflict-affected regions. They also suggested incorporating factors like population resilience, hybrid threats, and cognitive influences. A significant technical challenge mentioned was the difficulty of finding high-quality data proxies at the correct temporal and spatial scales. These insights suggest that while the current model is strong, integrating these complex socio-political layers would enhance its predictive power.

To conclude, the validation survey confirms that the machine learning framework captures real-world disaster drivers accurately. Both the statistical model and the expert group identified Population Exposure as the primary risk factor. Most participants endorsed the

data-driven approach for its objectivity and ability to reduce human bias. This significant overlap between computational results and expert opinion provides strong methodological triangulation. While the core findings are robust, practitioners highlighted the need to better account for governance and cascading crises in the future. This feedback validates the AGILE project's current results while offering a clear path for future model refinement.

Chapter 6 Conclusion

The completion of Task T2.3 within WP2 marks a significant advancement in the data-driven analysis of HILP events. A substantial historical database containing over 6000 disaster records from the 21st century was successfully analyzed using a dual-stage machine learning framework. The primary goal of identifying and classifying HILP events was achieved through the application of unsupervised learning techniques. While Self-Organizing Maps provided initial insights, the lack of explainability of the generated maps reduced their practical usefulness. Hence, the K-means clustering algorithm was adopted. It was found to be more effective for distinct event separation. This data-driven classification allowed for a clear differentiation between HILP and non-HILP disasters based on observed historical impacts. The algorithm labelled nearly 9% of the historical disasters as HILP events.

The second objective of analyzing factors that influence disaster escalation was addressed through a supervised machine learning model. The Random Forest algorithm successfully linked the classification of HILP events to underlying risk drivers, including hazard type, exposure levels, socio-economic vulnerabilities, and lack of coping capacities. The proportion of the population affected per million was identified as the most influential predictor in the model. Furthermore, social factors such as institutional trust and community engagement were found to be critical determinants of whether a hazard escalates into a catastrophic event. The importance of these socio-economic indicators emphasizes that HILP outcomes are emergent properties of vulnerable, interconnected systems. The outcomes of the supervised machine learning model led to the development of actionable guidelines and mitigation measures for HILP events.

The validation of these findings was conducted through extensive stakeholder engagement as outlined in the third objective. Internal verification was completed during the AGILE project's 4th General Assembly in Rotterdam, where partner feedback was used to refine the model's parameters. External validation was subsequently achieved at the International Climate Resilience Conference in Munich in October 2025, ensuring that the methodology and results remain relevant to the global practitioner community. Further, through an internal stakeholder survey, we validated the methodology and outcomes of T2.3.

In conclusion, the activities of Task T2.3 have been executed successfully and in full alignment with the project's Description of Action. A comprehensive link between historical impact patterns and systemic risk drivers was established. The integration of unsupervised and supervised machine learning has provided a rigorous explanatory framework for HILP

events. The resulting guidelines are intended to bolster institutional resilience and improve preparedness for future systemic shocks.

Chapter 7 References

1. Pescaroli, G., et al., *Definitions and Taxonomy for High Impact Low Probability (HILP) and Outlier Events.* International Journal of Disaster Risk Reduction, 2025: p. 105504.
2. De Groeve, T., K. Poljansek, and L. Vernaccini, *Index for risk management-INFORM.* JRC Science for Policy Reports (Brussels: European Commission), 2015.
3. Reduction, U.O.f.D.R., *The Human Cost of Disasters: An Overview of the Last 20 Years (2000–2019)*. 2020, UN Office for Disaster Risk Reduction Geneva, Switzerland.
4. Group, W.B. *Metadata Glossary Government Effectiveness: Estimate*. 2025 [cited 2025 July 16, 2025]; Available from: https://databank.worldbank.org/metadataglossary/worldwide-governance-indicators/series/GE.EST.
5. Surveys, I.V., *Justice system (courts) – Integrated Values Surveys" [dataset]. Integrated Values Surveys, "Integrated Values Surveys (IVS) Version 4*. 2024.
6. V-Dem, *Civil society participation index – V-Dem" [dataset]. V-Dem, "Democracy report v15" [original data]*. 2025, Our World in Data.
7. Alexander, D., *Disaster and crisis preparedness*. 2021, Oxford University Press.
8. Kohonen, T., *The self-organizing map.* Proceedings of the IEEE, 2002. **78**(9): p. 1464-1480.
9. Kohonen, T., *Essentials of the self-organizing map.* Neural networks, 2013. **37**: p. 52-65.
10. Sinaga, K.P. and M.-S. Yang, *Unsupervised K-means clustering algorithm.* IEEE access, 2020. **8**: p. 80716-80727.
11. Syakur, M.A., et al. *Integration k-means clustering method and elbow method for identification of the best customer profile cluster*. in *IOP conference series: materials science and engineering*. 2018. IOP Publishing.
12. Ezell, B.C., *Infrastructure vulnerability assessment model (I-VAM).* Risk Analysis: An International Journal, 2007. **27**(3): p. 571-583.
13. UNDP, *Human Development Report 2025.* UNDP (United Nations Development Programme), 2025.
14. Reduction, U.N.O.f.D.R. *Sendai Framework Terminology on Disaster Risk Reduction "Capacity"*. 2017 December 2, 2024]; Available from: https://www.undrr.org/terminology/capacity#:~:text=Coping%20capacity%20is%20the%20ability,during%20disasters%20or%20adverse%20conditions.
15. Coppedge, M., John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David, et al., *V-Dem Codebook v15*. 2025, Varieties of Democracy (V-Dem) Project.

16. OECD, *Foreign aid received – Official donors" [dataset]. OECD, "OECD Official Development Assistance (ODA) - DAC2A: Aid (ODA) disbursements to countries and regions*. 2025.
17. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
18. Vesanto, J. and E. Alhoniemi, *Clustering of the self-organizing map.* IEEE Transactions on neural networks, 2000. **11**(3): p. 586-600.

Appendix A

We developed a precedence variable as a representation of likelihood. To test this mechanism empirically, events with and without precedence were compared in terms of mortality and economic losses. Figure 21 shows the comparison. The horizontal axes show the presence or absence of precedence in binary terms as explained in Section 2.3. The vertical axes show the deaths and economic losses normalized with respect to the total population of the country and nominal gross domestic product of the country for the year in which the disaster occurred, respectively. Normalization neutralized the effects of population size and economy among the countries in the dataset.

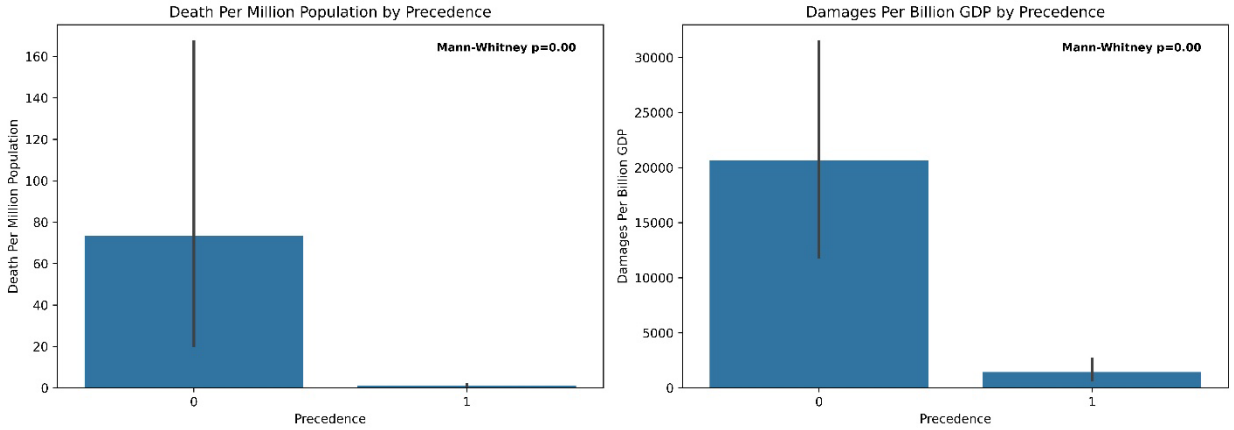


Figure 21 Effect of Precedence on Disaster Impact

The results revealed that events without a recorded precedence caused, on average, 73 fatalities per million people, while events with precedence caused only 1 fatality per million people. Similarly, events without precedence caused approximately $21000 economic loss per billion dollars of nominal GDP compared to just $1432 economic loss per billion dollars of nominal GDP for the events with a precedence. The differences are also statistically significant based on Mann-Whitney U test. The p-values for the Mann-Whitney U tests can be found on the top-right corner of Figure 21.

Appendix B

Due to the unbalanced nature of the minority class, oversampling was implemented on the training set. It should be noted that no oversampling was performed on the test set. SMOTE (Synthetic Minority Over-sampling Technique) was employed to address the severe class imbalance where HILP disasters represent only 8.7% of the dataset. Synthetic HILP samples are generated by identifying the k-nearest neighbors of each minority instance and creating new observations along the line segments joining them to their neighbors. This process is repeated until the training set achieves balance between HILP and non-HILP classes. Unlike simple duplication, SMOTE produces diverse synthetic examples that capture the local structure of the minority class in feature space, helping the Random Forest classifier learn more robust decision boundaries. Critically, oversampling was applied only to the training partition while the test set remained untouched to ensure unbiased evaluation of the model's true predictive performance on the original imbalanced distribution.

Appendix C

Validation of Data-Driven Drivers for HILP Disasters

We have developed a two-stage machine learning framework to uncover the drivers of High-Impact Low-Probability (HILP) events. We value your expert opinion to help validate our findings against real-world experience. This survey is anonymous and takes less than 5 minutes.

* Indicates required question

Section 1: Expert Profile

1. 1. Which best describes your primary area of expertise? *

Mark only one oval.

- ◯ Disaster Risk Management / Emergency Response
- ◯ Climate Science / Hazard Modelling
- ◯ Policy / Governance
- ◯ Socio-economic Analysis
- ◯ Other

2. 2. How many years of experience do you have in this field? *

Mark only one oval.

- ◯ 0-5 years
- ◯ 6-10 years
- ◯ 10+ years

Section 2: Validating the Drivers (Outcomes)

Context: Our machine learning model identified specific drivers for HILP events. We want to know if this matches your experience.

3. 3. To what extent do you agree with the following factors as critical drivers of HILP events? *

Mark only one oval per row.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Institutional Trust	◯	◯	◯	◯	◯
Community Engagement	◯	◯	◯	◯	◯
Population Exposure	◯	◯	◯	◯	◯
Hazard Type (Flood, Earthquake, Storm, etc.)	◯	◯	◯	◯	◯
Socio-economic Status (e.g. Unemployment)	◯	◯	◯	◯	◯

4. 4. If you could only pick ONE factor that most often turns a normal hazard into a HILP disaster, which would it be? *

Mark only one oval.

- ◯ Institutional Trust
- ◯ Community Engagement
- ◯ Population Exposure
- ◯ Hazard Type (Flood, Earthquake, Storm, etc.)
- ◯ Socio-economic Status
- ◯ Other
- ◯ Other: _____

Section 3: Validating the Methodology

5. 5. Do you consider a data-driven definition useful? Do you think it is more useful to identify HILP events using insights from historical data than from expert judgement? *

Mark only one oval.

- ◯ Yes, data driven approach is more useful.
- ◯ No, expert judgements are more practical.

Section 4: Open Feedback

6. 6. Is there a critical factor driving High-Impact disasters that you believe our data-driven model might have missed?