**A supervised learning classifier sensitive to false negatives built on multidimensional vulnerability features for cold-related disaster risk in Puno, Peru**

**Abstract**

This paper trains a household level disaster risk classifier for cold-related disasters based on supervised machine learning algorithms. The households’ features considered for this task come from multidimensional vulnerability to disasters accounting for economic, health, social and geographical vulnerability. In the geographic boundary of Puno case of study, there is an overwhelming exposure to cold-related disasters, thus vulnerability is hypothesized to explain disaster risk classification. Two models were tested: Logistic Regression and Random Forest Classifier. Hyperparameters of such models were optimized such as they maximize models’ ability to produce accurate predictions and to minimize false negatives that are the elements in the confusion matrix that produces deprivation costs. In the test dataset, Logistic Regression achieved a Matthews Correlation Coefficient of 63.19% and a Negative Predictive Value of 84.15%, while Random Forest Classifier achieved 62.47% and 85.71% respectively. After the application of the proposed algorithm, Negative Predictive Value, that controls the false negatives, increased without trading off model performance as is suggested by F1-Score, AUC-ROC and Accuracy metrics. Based on experiments with different sizes of test dataset, Random Forest Classifier outperformed Logistic Regression. Feature importance drawn from features contribution to reduction in entropy in the construction of the forest suggest that per capita expenditure, household localization in a rural area, altitude, access to public goods and concrete walls drove the disaster risk classification. Further research must propose strategies to externally validate the predictive model, and to analyze the causality of most important features regarding endogenous disaster risk classification.

# Introduction

Clime-related hazards are more frequent in this century than in the previous one (EM-DAT, 2022). This may be explained mainly by the increase in global warming and population sizes which, in turn, pressure on natural resources generating harmful outcomes for the environment (Keja-Kaereho & Tjizu, 2019). Disasters are not natural, as the same hazard would lead to different outcomes in different locations worldwide (Besiou et al., 2021). Disaster risk is the outcome of interactions of hazard, vulnerability, and exposure (UNDRR, 2015; Wright et al., 2020). The impact of disasters depends on the degree of vulnerability but also on the scale and magnitude of the hazard and the level of exposure. Hazards are phenomena that might harm humans, animals, and the environment, destroying a specific geographic position in a period (Preciado, 2015). Although hazards are mostly known to be an occurrence that human beings cannot control, human interaction with the environment has caused an increase in the frequency of clime-related hazards (Shabani 2022). Vulnerability shapes the damage that a natural hazard could cause and is entirely defined by anthropogenic conditions (Bolin, 2006). Exposure is the geographical conditioning of infrastructure, housing, and other tangible assets into hazard-prone areas (Mattea, 2019). Disaster risk is the outcome of interactions of hazard, vulnerability, and exposure (UNDRR, 2015; Wright et al., 2020).

Proactive disaster risk reduction is essential for communities affected by recurrent disasters. As suggested by Besiou et al. (2021), disaster risk management phases are not independent of each other. Thus, proactive disaster risk reduction activities that are carried out before a disastrous event could help to mitigate risks and create savings that communities may use for further development and building of resilience that is urgent due to the increasing magnitude and frequency of disasters. Puno is affected by recurrent cold waves and severe winter conditions. Peruvian’s South Andean Region is especially susceptible to these types of hazards. Since 2000, considering world-total historical data on disasters caused by Extreme Low Temperature Events (ELTEs) registered in EM-DAT, 21.28% of them have affected this geographic boundary. cordPuno is a rural and low-densely populated region located in the southeast of Peru. Puno can be considered the epicenter of ELTEs affecting PSAR, as 70.00% of events registered in EM-DAT affected Puno from 2003 to 2015. aryetc Cold waves and severe winter conditions or cold-related disasters that result from systematized and localized ELTEs are recurrent in Puno. That is why research on proactive disaster risk reduction would greatly impact this region.

This work focuses on disaster preparedness following a data-centric approach (EM-DAT, 2022). The main objective of this research is to predict which households would need to be prepared for a disaster that cold waves or severe winter conditions can trigger. This prediction must be accurate for the households that are at risk. When a predictive model misclassifies positive outcomes, defined as households at risk, deprivation costs are created (Gutjahr and Fischer, 2018 and Holguin-Veras et al., 2013); these cases are named false negatives. The model must give greater importance to accurate prediction of disaster risk, even if it implies that some households that do not have risk are being misclassified. Considering these objectives, the proposed methodology uses supervised learning algorithms, Logistic Regression and Random Forest Classifier, with data from Peruvian National Household Survey for Puno, 2019 to learn a binary classifier that discriminates which households are at risk of being affected by a cold-related disaster. Machine learning would help to build a risk screening tool that can be tuned, in terms of models’ hyperparameters, to maximize predictive power considering the importance of false negatives.

The proactive intervention on Puno may significantly impact the disaster response and recovery. Following Holguin-Veras et al. (2013), resources invested in response and recovery includes logistic costs and deprivation costs. An optimized predictive model would identify which households would be the target of proactive interventions. Puno is a case of study characterized by spatially dispersed final demand points and high peaks of deprivations caused by accumulated vulnerabilities (Kim and Sohn, 2018; Quiliche et al., 2021), accurate forecasts are of special importance. Assessment of delivery strategies, transportation costs and their balance with deprivation costs are left for future research as the objective function is of main concern of humanitarian logistics.

The contribution of this paper is twofold: [1] introduce vulnerability-based disaster risk prediction, in contrast with other predictive strategies based on meteorological, geophysical or geographical modelling and [2] propose a hyperparameter optimization algorithm based on domain requirements such as minimization of false negatives. The key element for hyperparameter optimization procedure is the confusion matrix of the predictive models, as logistics costs depend on False Positives and True Positives (which represent demand points that must be attended), True Negatives means that no delivery is required, and deprivation costs arises from False Negatives (which represent demand points that need essential supplies, but the model misclassify their risks so aid goods are not being supplied). The experimental setting for hyperparameters’ optimization will consider confusion matrix metrics by co-optimizing on Matthews Correlation Coefficient (MCC) and Negative Predictive Value (NPV). HPO is usually based on one metric, but the proposal includes sequential optimization of MCC and NPV, where maximization of MCC aims to minimize social costs and maximization of NPV aims to minimize deprivation costs.

The learned predictive model is expected to contribute to reduce social costs while considering the importance of deprivation costs Holguin-Veras et al. (2013). As the focus is on disaster preparedness, the predictive model will be used to identify the final demand points that need pre-positioning of supplies, thus producing information regarding the number of supplies required or the demand for humanitarian aid to perform proactive interventions. In the context of disastrous events, the value of information regarding where to pre-position supplies and how many supplies to pre-position is high, as those supplies aim to reduce the expected damages on households’ livelihoods that are strongly linked to agriculture and livestock (Quiliche and Mancilla, 2021).

In this context, the remaining of this paper is divided into five sections. Section 2 describes the main work on SLAs, so as machine learning applications to disaster risk management and emergencies assessment. Section 3 details the data collection methods and processing pipeline, the experimental setting will be also defined in this section. Section 4 brings the results regarding hyperparameter optimization and deprivation costs will be presented. In the discussion section, the main results will be debated with the scope of humanitarian logistics principles. Finally, in the conclusions section, the article will be closed with a set of recommendations for improvements on disaster preparedness strategies and future research avenues, including external validation strategies of the proposal.

# Theoretical framework

## Impacts of disasters triggered by ELTEs in Puno, Perú

In this paper, the main concern is about the impact of ELTEs on the final echelon of the Humanitarian Supply Chain (HSC) that is the household at risk of being affected by a disaster. After a disaster is triggered, in the response phase, supplies are delivered to affected communities (Alexander, 2002; Ferreira, 2012) following a wide variety of strategies that may vary according to the specific characteristics of disasters (Apte and Yoho, 2011). At the final echelon of the HSC, disasters affect households and their inhabitants. In this case of study, considering that 66.54% of households are settled in geographic strata with less than 500 inhabitants, rural communities are being affected by ELTEs in a particular way that is defined by their local livelihoods (Quiliche and Mancilla, 2021). The main impacts of disasters triggered by ELTEs on rural communities implies the destruction of crops, livestock and, in the worst scenario, it causes losses of human lives. For urban settled households, that represent the minority of cases in the sample, the expected losses may be lower because of difference in infrastructure, and differences in their livelihoods (López-Bueno et al., 2021). The causal flow from ELTEs to Disastrous events that affects households is summarized in the following Figure 1:

**Figure 1.** Causes of cold-related disastrous events affecting households in Puno

Graphical user interface, application

Description automatically generated

Regarding mortality, López-Bueno et al. (2021) performed statistical analysis of mortality rates caused by cold waves in both urban and rural areas of Madrid, Spain. The authors conclude that the main risk drivers of mortality rates are socioeconomic, they estimate an index of socioeconomic deprivation that is positively related to mortality rates, controlling for differences between urban and rural municipalities. Amirkhani et al. (2022) found an interesting pattern for a cross-section of countries around the world for the period 1999-2018 using EM-DAT (2022): cold waves and severe winter conditions produced more deaths on middle-income countries than in high-income ones and, for all cases, CO2 emissions are strongly correlated with both frequency of cold waves and overall temperature variability. Regarding the livelihoods of inhabitants in Peru, Quiliche and Mancilla (2021) stated that rural households make the decision to diversify their income sources (coming from crops, livestock, among other by-products) considering the risk of not being able to guarantee their own subsistence and the reposition of their livelihoods. Rural households must maintain a minimum level of food production, reposition and having a monetary surplus to exchange for health and education services in local markets in contexts of severe deprivations and ELTEs for the case of Puno.

## Complexities arising from deprivation costs in Puno, Perú

In contrast to commercial logistics, humanitarian logistics deals with HSCs and does not count for logistic costs minimization alone, but also for minimization of human suffering (Van Wassenhove, 2006 and Tomasini and Van Wassenhove, 2009). In this line, reducing human suffering can be challenging due to ambiguous objectives(Tomasini and Van Wassenhove, 2009)**.** For instance, Holguin-Veras et al. (2013) were concerned about deprivation costs that are defined as the cost associated with human suffering from the lack of access to essential supplies in the context of a disaster aftermath, when peaks of demand are observed. According to review carried out by Shao et al. (2020), it is common to find a trade-off between logistics costs and deprivation costs, being social costs the sum of logistics costs and deprivation costs. In this regard, this paper referrers to the experimental results of Gutjahr and Fischer (2018):

“We are concerned with the decision on the frequency of periodical relief commodity deliveries to demand points in the response phase after a disaster. In some disaster scenarios, the accessibility of demand points and therefore also the supply costs heavily vary. This has often the effect that not all demand points are provided with relief commodities with the same frequency. Thereby, an equity issue is raised, which makes the problem relevant for our investigation” (Gutjahr and Fischer, 2018).

The authors use data from Nepal earthquake 2015 to show, by using a simplified mathematical model, that practically irrelevant final reductions of average deprivation costs result in substantial increases of equity in the optimal solution between different demand points. In the case of Nepal earthquake 2015, Gutjahr and Fischer (2018) showed that equity in aid distribution can be achieved at a relatively small increase of logistics costs, and thus reaching an equilibrium between two objectives: deprivation and logistic costs. Characteristics of disasters triggered by ELTEs in Puno, Peru reproduces this trade-off (see Figure 2) under the following conditions for humanitarian operations due to the large spatial dispersion of final demand points that are households at risk of being affected by ELTEs: the human and economic resources for the response are scarce as Puno is a poor region, the aid must be delivered as soon as possible after a disastrous event to prevent losses, and the operations must seek to meet the maximum demand possible to guarantee equity in the optimal solution.

In this line, the two identified objectives, logistic and deprivation costs, are formally defined in the following Equations (1.1. to 1.3.):

(1.1.)

(1.2.)

(1.3.)

The trade-off between and lays in the following decision-making concerns: the final decision is sensible to the level of dispersion of final demand points given any configuration of cost parameters (Gutjahr and Fischer, 2018; Chong et al., 2019). Excluding some demand points from the response operations can reduce a lot the complexity and the total cost of the response. The challenge is to co-optimize and to optimize the so called that matters because communities need to balance their budget to invest in resilience to be better prepared for future disastrous events related to ELTEs. The optimization of logistic costs alone () would cause breaks with impartiality principle of humanitarian logistics (Tomasini and Van Wassenhove, 2009). In this line, Leiras et al. (2017) stated that impartiality in humanitarian logistics must seek to reach an equilibrium between ambiguous objectives.

**Figure 2.** Spatial distribution of households exposed to ELTEs

Map

Description automatically generated

On the one hand, spatial dispersion of final demand points necessarily increases logistics costs () during the response phase (Balcik and Beamon, 2008; Ferreira, 2012). On the other hand, deprivation costs (), that are related to the level of socioeconomic vulnerability or socioeconomic deprivation (Villarroel-Lamb, 2020; Wang et al., 2020; Szczyrba et al., 2021), tend to be higher for households that are settled far from cities, because they are already deprived from consumption of essential goods and services in the business-as-usual scenario (Quiliche and Mancilla, 2021). In general, when a disaster is triggered, households that are settled far from cities suffer from peaks of deprivations due to delays in distribution in a context of sudden increases of demand due to disastrous events (Shao et al., 2020). The following Table 1 illustrates three logistic-related outcomes arising from the trade-off between and for Puno:

**Table 1.** Possible outcomes considering deprivation costs

|  |  |  |  |
| --- | --- | --- | --- |
| **Logistic cost** | **Low cost** | **Equilibrium** | **High cost** |
| **Solution** | Logistic costs optimized without considering deprivation costs | Logistic costs and deprivation costs co-optimized | Deprivation costs minimized without considering logistic costs |
| **Outcomes for Puno, Peru** | Critical demand points are not supplied with enough aid | Critical demand points are optimally supplied at a balanced logistic cost (achieving equilibrium in social costs) | Critical demand points are fully supplied at a high logistic cost (social cost is high due to high logistic costs) |

Authors own elaboration

## Characteristics of disasters triggered by ELTEs in Puno, Peru

Following EM-DAT (2022) taxonomy of disasters based on Integrated Research on Disaster Risk (IRDR, 2014), ELTEs produces disasters caused by cold waves and severe winter conditions. The data from EM-DAT (2022) registers four disastrous events that have affected Puno in the period 2003-2022 (Table 2):

**Table 2.** Possible outcomes considering deprivation costs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification: extreme temperature** | **Year** | **Magnitude** | **Start** | **Duration** | **Total Affected** |
| Cold wave | 2003\* | -28 °C | July | 38 days | 1839888 |
| Cold wave | 2004 | -35 °C | June | 30 days | 2137467 |
| Severe winter conditions | 2007 | -20 °C | April | 90 days | 884572 |
| Cold wave | 2015 | -20 °C | May | 141 days | 200620 |

Authors own elaboration. (\*) represents the only registered case that included an official response from OFDA.

* These disastrous events are dispersed and sudden onset at the region of impact (Van Wassenhove, 2006; Apte and Yoho, 2011), which means that they affect a large scale of territory and produce dispersed peaks of human suffering that are magnified by socioeconomic vulnerability (Villarroel-Lamb, 2020; Wang et al., 2020; Szczyrba et al., 2021).

The following Figure 3 show the evolution of minimum temperature recorded by the SENAHMI weather stations (2022) that are located within Puno’s territory. The daily average of minimum registered temperatures among the stations for the period 2009-2012 is reported. This provides an estimate on the level and variability (standard error) of minimum temperature that can be matched with the characteristics of cold waves and severe winter conditions to explore the nature of their seasonality. According to EM-DAT (2022), the disastrous events of 2007 and 2015 were triggered at the beginning of April and May, and lasted until July and September, respectively. In the SENAHMI (2022) dataset, seasonality is observed in terms of average minimum temperature for these months, the temperature drop tends to start in April and last until July, and then slowly return to normal levels in September.

**Figure 3.** Time series plot for average minimum temperature in Puno 2009-2012

|  |  |
| --- | --- |
| Chart  Description automatically generated | Chart  Description automatically generated |
| Chart  Description automatically generated | Chart  Description automatically generated |

According to an institutional report from published by Food and Agriculture Organization (Alarcón and Trebejo, 2010) based on data from SENAHMI 76.2% of the territory is above 3500 meters above the sea level and had minimum temperatures on the range from -16 °C to 8 °C for an average of 15 days for June, July and August. In this report, the authors conclude that during the period 1969-2010, for each year there was at least one cold wave[[1]](#footnote-1), this does not mean that for each instance a disastrous event was triggered. The historical data about disastrous events is limited, but the report states that the hazards are seasonal and recurrent. This fact characterizes disaster risk for households settled over the Puno region: the probability that ELTEs, such as cold waves or severe winter conditions, will trigger disastrous events on population is considerably high, despite the underreporting of these types of disaster found in EM-DAT (2022) for low-income countries (Amirkhani et al., 2022). Humanitarian operations must seek to create proactive interventions to mitigate the losses (Van Wassenhove, 2006; Bosher et al., 2021).

## Vulnerability against cold waves and severe winter conditions

Vulnerability is a multidimensional concept based not only in economic scarcity, but also in lack of access to basic services, lack of health, education, low social development (Pessoa, 2012) and geographical exposure for the case of disasters (Ullah et al., 2021). Several studies conceptualized vulnerability against clime-related disasters as the quality or state of being exposed to the possibility of being harmed by a disaster (Christian et al., 2021; Sahana et al., 2019; Tasnuva et al., 2020; Ullah et al., 2021). This possibility increases when a set of characteristics are met. For the case of the households settled in Puno and according to available data a multidimensional framework is proposed in Figure 4.

**Figure 4.** Dimensions of vulnerability against clime-related disasters

Diagram, text

Description automatically generated

According to Figure 6, we propose that vulnerability has four dimensions: economic, health, social and geographical. Low income and bad infrastructure are the main drivers of vulnerability to disasters according to Tasnuva et al. (2020). Bad outcomes in health, such as a high prevalence of chronic illness could be related with a higher vulnerability (Djalante et al., 2020). Certain configurations of socio-economic variables make households especially vulnerable, such as unemployment, and low educational achievement, there is evidence that younger and female head of households is related to the probability of being affected by a disaster (Rapeli, 2017). Geographical vulnerability depends on household location, which at the same time is determined by economic vulnerability: households located in vulnerable areas tend to be poor and this magnifies the vulnerability condition (Mattea, 2019).

It is important to differentiate vulnerability from risk. According to UNDRR (2015) and Twigg (2004), in the context of disasters, risk is the combination of vulnerability, hazard and exposure, where vulnerability is defined as the foreseeable consequences of a damaging event on entities such as human lives, health, wealth, or environment. Hazard is the loss or injury caused by a potentially damaging physical event. Exposure is the property, people, systems, or other elements present in hazard zones. Vulnerability is operationalized as the human factor of disaster risk that shapes disaster risk. The predictive analysis would be entirely based on vulnerability factors. The geographical variables could capture a certain level of exposure, but it is important to notice that other variables could be also correlated with exposure. This paper contributes with an empirical application and evidence based on predictive analytics and machine learning applications. In the disaster risk assessment literature empirical evidence from data-mining methods is missing according to Behl and Dutta (2018).

## Using data science to improve humanitarian operations in Puno, Peru

The Sendai Framework for disaster risk reduction proposes to reduce disaster mortality, reduce direct economic losses (Mors, 2010; Lu et al., 2021) and to invest in disaster risk reduction for resilience (Wright et al., 2020). Regarding such goals, machine learning applications in the field of DRM, contributed with the improvement of the monitoring of information during emergent situations and decision-making under time-sensitive conditions (Lu et al., 2021). Humanitarian operations are naturally coupled with chaotic environments (Tomasini and Van Wassenhove, 2009), information processing and data-driven decision making is key to transform information into insights that could help to solve common challenges in humanitarian operations such as integration between stakeholders of the HSC (Balcik et al., 2010; Leiras et al., 2018; Sokat et al., 2018). Integration between actors in humanitarian response respond to data-driven systems that help in decision-making. Some inefficiencies that arise from lack of integration are, for example, material convergence (Holguin-Veras et al., 2014) and transport and distribution bottlenecks (Alcántara-Ayala, 2019). According to Behl and Dutta (2018)’s review conducted to identify current challenges in management of HSCs, integration between stakeholders is a complex task that requires sophisticated solutions. To deal with integration in humanitarian logistics, several authors recommend the use of data science methodologies (Fayyad and Shapiro, 1996) that includes all the methods based on structured or unstructured data to extract knowledge in the form of patterns to solve real problems (Behl and Dutta, 2018; Fernández-Luque et al., 2018; Sokat et al., 2018; Lu et al., 2021). In this regard, this paper contributes to the branch of the literature on humanitarian operations that aims to embed data analytics and Machine Learning-based predictive analytics to decision-making in humanitarian operations:

“With the advent of big data, cloud computing and developed technologies like drones, it would be interesting to look at algorithmic models to standardize the process of handling post-disaster operations” (Behl and Dutta, 2018).

On the other hand, considering that this case implies seasonal hazards that produce recurrent disasters in Puno, human suffering alleviation and improvement in local livelihoods are valuable to save resources and consider the future availability of budget to deal with future disastrous events (Bosher et al., 2021). Humanitarian logistics are embedded in each one of DRM phases depicted in the disaster management lifecycle: risk mitigation, disaster preparedness, response, and recovery (Petak, 1985; Van Wassenhove, 2006). However, there is scarce literature that aims to optimize the entire disaster management lifecycle (Bosher et al., 2021) or at least impact over all the phases involved in DRM (Alexander, 2002; Ferreira, 2012). In the same line, Bosher et al. (2021) points out that:

* The disaster management lifecycle is presented as a closed loop that prevents humanitarian logisticians to move through the phases and include new inputs and activities that may lead to no further disasters when disasters are recurrent
* Too much importance is given to the disaster event itself. As a result, most activities are biased towards emergency management or disaster response, which can be illustrated in the results of systematic literature review done by Overstreet et al. (2011). This approach is defeatist and inaccurate and have caused obstacles integrating DRM phases to reach a new status-quo where disasters effects on communities are really mitigated (Wright et al., 2020).
* Considering temporal and resource considerations (Contreras, 2016).
* Considering identifying the underlying root causes of vulnerability drivers of disaster risk (Wisner and Lavell, 2017; Van Riet, 2021; Renteria et al., 2021).
* Acknowledge the role of complex systems and the ability to change (Coetzee and Van Niekerk, 2018).

This paper pretends to use data to train a ML model that is capable of classifying households as risk-prone in Puno. Assuming that the training is performed with success, the model would be able to predict if a household is going to need supplies of aid, and thus help in decision-making for pre-disaster risk reduction and planning activities (Bosher et al., 2021) considering the importance of deprivation costs in the training process, which represents a novel contribution on ML research, considering that HPO is non-trivial. The second contribution of this paper is that it proposes to integrate domain knowledge to the HPO process by co-optimizing deprivation costs and logistic costs (Amaral et al., 2021).

The following Figure 5 summarizes the trend in literature that uses ML models, algorithms, and even automated decision-making systems to improve humanitarian operations. The shape of the circles shows the estimated frequency of related studies based on exhaustive analysis of Lu et al. (2021) and literature review that was carried out. Every circle contains a topic and a representative paper in each topic. Supervised learning methods were used to perform statistical and predictive analysis to identify underlying characteristics that determines future recovery paths (Nejat and Gosh, 2016), generating data to map disaster risk for mitigation (Shafapourtehrany et al., 2022), deep learning-based image real-time detection of disasters (Webster, 2017) and rapid algorithms for decision-making in disaster response (Yan et al., 2021). Morss (2010) evaluates the impact of ML based predictions into real time communities’ interactions, decision-making and outcomes through the scope of disaster preparedness. There are some other papers that addresses more than one disaster phase at a time, but for simplicity those papers are not reported in this review.

**Figure 5.** Findings from literature review

Shape

Description automatically generated

ML applications focus on predictive analytics; thus, prediction is commonly used in literature to propose data-driven frameworks in risk mitigation and disaster preparedness phases of disaster management lifecycle (Ferreira, 2012). Bosher et al. (2021) refers to these phases as proactive Pre-Disaster Risk Reduction and Preparedness Activities (PDRRPA). The third contribution of this paper is that, through applied data science methods, the Disaster Management Helix, which is a graphic representation of disaster management lifecycle that aims to conceptualize the evolution of DRM through time in a community, can be optimized as whole. As is shown in Figure 7, The Disaster Management Helix for the case of Puno, reproduces the following pattern:

1. High impact of first disastrous event, that has also high probability of being seasonally repeated.
2. Resources and efforts invested on response and recovery in the disaster aftermath are considerably high and disproportionate.
3. The budget for following PDRDPA is substantially lower, and this induces a cyclical scheme in which it turns particularly hard to deal with future disasters.

**Figure 6.** Theorical representation Disaster Management Helix optimization

Chart, line chart, histogram

Description automatically generated

In specific, given that risk mitigation implies policies that aim at reducing disaster vulnerability in the long run, thus requiring more resources, the proposal of our approach is to focus on disaster preparedness to change the status-quo of DRM from ‘non-optimized helix’ to ‘optimized helix’. The implementation of predictive models that are capable of decide which household must be supplied with aid would lead to an optimization of resources through several mechanisms such as stock pre-positioning (), proactive delivery, and gradual delivery (Apte and Yoho, 2011). The better utilization of scarce resources would increase the budget that will be assigned to response and recovery activities in future disasters, and as Bosher (2021) suggest, this might lead to a future scenario where:

“Informed decision making that results in desirable social, physical, economic and environmental development pathways that avoid disaster risk creation” (Bosher et al., 2021).

For the case of Puno, social costs are being minimized by optimizing preparedness strategies to reduce economic losses and resource utilization for response and recovery over the time in communities affected by cold waves and severe winter conditions.

# Materials and methods

## Data collection methods and the classification problem

Raw data on households’ disaster vulnerability dimensions were collected from the National Household Survey (NHS) that was carried out by Peruvian’s National Institute of Statistics and Informatics in 2019. Over the survey modules: population and housing (modules 100 and 200), education (module 300), health (module 400), employment (module 500) and democracy and transparency (module 612) were selected, according to insights pointed in the literature on possible features that may predict disaster risk classification and thus output a decision regarding aid delivery (UNDRR, 2015; Salazar-Briones et al., 2020 and Renteria et al., 2021). In the survey, questions are asked to the head of the household and 23.33% of them are answered by another informant.

The learning target is a binary indicator as shown by Equation 1. The empirical classification problem will be assessed through the lens of supervised learning techniques. Regarding classification classes, module 612 asks the following question for each household: in the last 12 months, your house has been affected by natural disasters (drought, storm, plague, flood, etc.)? One identified problem is that such a question does not provide specific information about the type of disaster associated to the referred risk. To argue on the reliability of the use of this variable to measure households’ risk of being affected by a disaster triggered by cold waves or severe winter conditions without applying any feature engineering method to it we consider the following facts:

1. For the specific case of Puno, there is an overwhelming prevalence of risks related to low temperatures[[2]](#footnote-2). Furthermore, in this area lays 99.1% of the total population of Puno.
2. The average household’s monthly earnings are S/. 470.2 and the poverty line is estimated on S/. 352. Based on data from NHS 48.6% of households are poor for year 2019 and thus risk-prone because of their economic vulnerabilities.
3. Rentería et al. (2021) found a strong statistical correlation between risk classification for different disaster types that support the hypothesis of ‘similar vulnerability dynamics between disasters. This means that, at the household level, if a household is at risk of being affected by floods, then is very likely that it is also affected by landslides. The mechanism that explains this correlation is the vulnerability conditions shared by these households.

Considering this evidence, it seems reasonable to operationalize the target variable as in Equation 1: equal to one when the household is at risk of being affected by a disaster triggered by cold waves or severe winter conditions. Supervised learning will use this variable as the target for the classification problem.

**Equation 1.** Binary variable measuring household disaster risk

## Dataset particularities and pre-processing

The feature space extracted from NHS is multidimensional in the sense that several variables were collected for vulnerability dimensions: economic, health, social and geographical. This overcomes empirically the over-simplification of disaster vulnerability that suppose that vulnerability is socioeconomic and ignore the dependence of the other factors (Villarroel-Lamb, 2020; Regal, 2021; Szczyrba et al., 2021 are some examples). The proposed approach entails greater empirical complexity as a greater number of features are considered for the model training process:

“High-dimensional datasets bring a lot of information to people, at the same time, because of its sparse and redundancy, it also brings great challenges to data mining and pattern recognition” (Xuan et al., 2019).

From NHS (2019), 86 household-specific features were collected. A greater number of features will increase the computational time required for HPO for each supervised learning algorithm. As we are dealing with a wide set of categorical features, dimensionality reduction techniques based on projection may not perform well and this may affect the predictive power of supervised algorithms[[3]](#footnote-3). To overcome this obstacle, we have selected supervised algorithms that incorporate feature selection in the training process. This training method is called sparse learning. By incorporating feature selection into the training stage, algorithms remove large amounts of redundancy and noise, and keep a subset of input features by maximizing predictive power (Xuan et al., 2019). Instead of generating information loss, that is the case for dimensionality reduction methods based on projection, sparse learning includes features’ regularization terms into their objective functions and reach maximum predictive power through balancing bias-variance trade off using cross-validation method to compare post-processing metrics that measure classification performance (Jian et al., 2008; Robert, 2011).

It is worth mentioning that the training process will be performed on Python 3.10 using Scikit-Learn 1.1.1 package. We propose the following sparse learning algorithms: Random Forest Classifier, XGBoost, Support Vector Classifier and Elastic-Net Logistic Regression. Following documentation guidelines, the training process of these classification algorithms speed-up when input features are in the same scale. Furthermore, scaling features to the same range improves interpretability of results regarding ‘feature importances’ that is a core element of sparse learning. Sparse-learning classifiers, as well as ensemble-based, can be trained successfully even in the presence of ‘dummy variable trap’ that turns classical statistical learners unfeasible.

* 1. Proposed supervised learning algorithms

In the comprehensive review of Zebari et al. (2020), the authors identified that ensemble methods could be useful for feature selection as they provide robust measures of features’ importance that are based on the likelihood of a feature to be able to predict the outcome. For example, Xin and Ren (2022) plots the contributions to outcome prediction for each tree in their Random Forest Classifier (RFC). Novel approaches such as the one of Anitha and Vanitha (2022) uses Extreme Gradient Boosting (XGBoost) to perform feature selection and use the output feature space for Stochastic Gradient Descent (SGD) training, this pipeline led the authors to a significant improvement in classification accuracy. There are infinite combinations of pipelines and each one of them could lead the sparse learning to different levels of scores regarding classification performance. Seminal work from Giovanelli et al. (2021) introduced an automatized framework for HPO in classification algorithms called AutoML, however the authors also states that a proficient data scientist with enough domain expertise may be able to outperform the algorithm and find better pipelines. Considering discussion above, Equations 2-5 will describe the selected sparse learning algorithms for classification:

**Random Forest classifier**

Often referred as CART algorithm (Jackins et al., 2021), RFC are composed of multiple decision trees that are pruned and then averaged to balance the bias-variance trade-off and maximize the predictive power of the ensemble. To train RFCs, the following steps must be followed (Xin and Ren, 2022):

1. Randomly select ‘n’ features from total ‘k’ features.
2. Randomly select ‘max\_samples’ number of samples with bootstrap method.
3. Among the ‘n’ features, calculate the first node using the best split point with Gini or Entropy ‘criterion’, following the rules defined by the parameters for each split:
   1. ‘Min\_samples\_split’: minimum number of data points placed in a node before the node is split.
   2. ‘Min\_samples\_leaf’: minimum number of data points allowed in a leaf node.
4. Categorize the node into daughter nodes using the best split with Gini or Entropy criterion
5. Categorize more daughter nodes until the tree reaches the depth equal to ‘max\_depth’
6. Repeat 1 to 5 steps several times equal to ‘n\_estimators’ to build such number of trees, which refers to the size of the forest.
7. Build the prediction algorithm by averaging the probabilistic prediction among the entire forest.

**ENLR**

Zou and Hastie (2005) proposed for the first time the Elastic-Net regularization technique, as a combination of Least Absolute Shrinkage Selection Operator (LASSO) and Ridge regularization terms. The adaptation to Logistic Regression was proposed in literature using different solvers and formulations, but the one that is used here is based on Pedregosa et al. (2011). The objective function is stated as follows:

Where is a data vector corresponding to observation , is the respective observation point for target classes.

**Support Vector Classifier (SVC)**

The SVC is a model based on the construction of Support Vector Machines (SVM) that are in essence hyper-planes that split the dataset based on their patterns following a target. The maximization problem aims to find the hyper-plane that maximizes the distance between feature space considering the target classes. The SVC first creates a linear separating hyper-plane and then uses kernels to project nonlinear data into a form that is linearly distinguishable (Shafapourtehrany et al., 2022). Once trained, the classifier predicts classes according to the following decision function:

Where, . Sparse learning applied to SVC requires to add a L2 regularization term to the objective function (L1 regularization is not available in current scikit-learn version for python). Although objective function is not described here, is important to state that this modification is needed to regularize parameters to reach sparse results on coefficients . The following kernel function is called Radial Basis Function (RBF) that is selected as the preferred kernel function as it is good at discovering non-linear hyper-planes target classes:

* 1. Post-processing metrics

The following Figure 2 shows the confusion matrix that illustrates performance of classification algorithms. The mostly used heuristic is to maximize the diagonals or the accuracy of the classifier. However, given the complexities described in Section 2.2, the classification problem demands a different approach. To describe such approach, the relationship between classifier performance metrics and logistic, deprivation and social costs (Holguin-Veras et al, 2013; Shao et al., 2020) will be next defined:

Figure 7

**Interface gráfica do usuário, Aplicativo, PowerPoint

Descrição gerada automaticamente**

**Table 3: Interpretation of confusion matrix elements**

|  |  |
| --- | --- |
| **Quadrant** | **Relation with social costs** |
|  | Cases where households do not have risk and the model classifies it correctly, so the model decides that they do not need supplies. Greater values in this quadrant save social costs, as no supplies are required by households that have no risk. The no-risk classification depends on a threshold for predicted probabilities, set to 50% by default. |
|  | Cases where households do not have risk and the model misclassifies them and decides that must be delivered with aid, thus generating undesired logistic costs (). |
|  | Cases where households have risk and the model misclassifies them and decides that they do not need supplies, thus directly generating deprivation costs on demand points that are not being supplied with aid when they need it. Following Section 2.5, deprivation costs must be emphasized, as the reproduce vulnerabilities. Furthermore, if ignored, peaks of deprivations may lead communities to peaks of resources utilization. In extreme cases, international help is required to cover demand from peaks of deprivations. |
|  | Cases where households have risk and the model decides that they must be delivered with aid, thus generating justified logistic costs () |

**Area Under the ROC Curve (AUC)**

This metric represents the distance between ‘no discrimination’ classifier (worst classifier that distributes uniformly the predictions over classes for any probability threshold) and tested classifier. It is defined in function of and coordinates at various probability threshold settings. The range of this metric varies in the closed interval so better classifiers are found when .

**Accuracy**

The estimation of accuracy represents the application of common heuristic where diagonal of confusion-matrix is maximized. The formula is given by . The range of this metric varies in the closed interval so better classifiers are found when .

**F1-Score**

Is defined as the harmonic mean of the and . The formula is given by . The range of this metric varies in the closed interval so better classifiers are found when .

**Matthews Correlation Coefficient**

This metric is in essence a correlation coefficient that lays in the [-1,1] interval. The formula is given by . It was selected to choose the best classifier as it tends to co-optimize all elements of the confusion-matrix for binary classifications (Luque et al., 2019; Chicco and Jurman, 2020). By maximizing this metric, the classifier is minimizing both deprivation costs and logistic costs.

**Negative Predictive Value**

This metric shows the performance of the classifier regarding negative classes. In the stated problem, negative classes are highly related to deprivation costs, thus misclassifying no-risk households may lead to peaks of deprivation (see Table 4) and other consequences described in Section 2.2. The formula is given by . The proposed HPO strategy will try to co-optimize MCC and NPV.

* 1. Experimental setting: a domain-based approach to HPO

Although there is no novelty in the use of such algorithms, nor in the training metrics, the novelty is on the constructed strategy for HPO, which will be further explained. Considering that the main justification to use a machine-learning approach here is to use the predictive models to support decision-making, as explained in Section 2.2, we define next the proposed algorithm for HPO:

Pseudo-algorithm

1. Define the space of the search of hyperparameters for each supervised algorithm.
2. Set cross validation method to repeated stratified cross-validation with ‘k=10’ folds and ‘r’ repeats.
3. Random search cross-validation with ‘n\_iterations=2000’. Estimate average AUC, Accuracy, F1-Score, MCC and NPV over folds and repeats.
4. Keep the ‘percentile=5’ best hyperparameter configurations based on average MCC.
5. Using the results above, select the hyperparameter configuration that maximizes NPV.

In Step 1, the space of search for HPO is defined. In Step 2, cross-validation strategy to shuffle data into train-test splits is selected as repeated stratified cross-validation, which is a useful method to reach robust solution in classification problems, as it returns stratified folds, where each fold contains the same proportions of samples of each target class as in the complete dataset. In Step 3, the strategy is to test 2000 random combinations of hyperparameters and estimate post-processing metrics for each experiment. However, the average metrics across ‘k=10’ folds and ‘r=2’ repeats will be used for further steps. In Step 4, keep the 100 better hyperparameters’ configurations (which is equivalent to ‘percentile=5’), by MCC. This step aims to get a higher NPV at a cost of small reduction of MCC to reduce potential deprivation costs that may arise by model predictions as they discriminate whether the household will be delivered with aid or not. In this case, the decision-making is concerned with disaster preparedness strategies, so if the model decides that a household must be delivered with aid, prior to disastrous event, it must be targeted in the preparedness planning.

Table 4: Parameter grid for supervised learning algorithms

|  |  |
| --- | --- |
| Supervised learning algorithm | Parameter distribution |
| Random Forest Classifier |  |
| Elastic-Net Logistic Regression |  |
| Extreme Gradient Boosting (XGBoost) |  |
| Support Vector Classifier (SVC) | C=[] |

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

# Results

The following Table shows descriptive statistics for categorical features (dummy-encoded) and the Table for numerical features. Additional pre-processing techniques were applied, in this case, categorical features with a frequency lower than 2% of samples were discarded in order to improve the results of supervised learning algorithms. Small frequencies in categorical features led to null models in the train-test split phase of the training process. As 10 folds were selected for cross-validation, the train-test split procedure entails a high probability of produce a split with a categorical feature equal to zero, which is the same as not considering it at all. Statistical analysis is recommended to investigate the importance of such features as they could be important, regarding disaster risk, or they could be noise.

## Descriptive statistics

Figure 8: Features' correlation heatmapGráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente

Table 5. Descriptive statistics

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Variable** | **Count** | **Percent** |
| Household exterior and access to public goods | Household with inlaid walls | 194 | 17.54% |
| Household with painted walls | 156 | 14.10% |
| Outside tracks are paved | 259 | 23.42% |
| Outside tracks are terrain | 311 | 28.12% |
| Outside paths | 226 | 20.43% |
| Lighting pole | 442 | 39.96% |
| No public good | 442 | 39.96% |
| Ownership and physical characteristics | Independent house | 944 | 85.35% |
| Household is a house | 955 | 86.35% |
| Household is totally owned | 913 | 82.55% |
| Tittle of ownership | 231 | 20.89% |
| Concrete walls | 287 | 25.95% |
| Concrete floor | 361 | 32.64% |
| Concrete roof | 228 | 20.61% |
| Overcrowded bedrooms | 374 | 33.82% |
| No other rooms than bedrooms | 124 | 11.21% |
| Access and use of basic services | Water network | 382 | 34.54% |
| Potable water | 531 | 48.01% |
| Quality water (chlorine) | 122 | 11.03% |
| Daily access to water | 668 | 60.40% |
| Drainage network | 382 | 34.54% |
| Electric lighting | 1017 | 91.95% |
| Candle lighting | 51 | 4.61% |
| Other lighting | 49 | 4.43% |
| GLP cooking | 485 | 43.85% |
| Wood cooking | 61 | 5.52% |
| Other cooking | 141 | 12.75% |
| Manure cooking | 417 | 37.70% |
| Phone | 34 | 3.07% |
| Cellphone | 918 | 83.00% |
| Cable TV | 114 | 10.31% |
| Internet | 142 | 12.84% |
| Household assets | Radio | 890 | 80.47% |
| Color TV | 536 | 48.46% |
| Black-White TV | 132 | 11.93% |
| Sound equipment | 92 | 8.32% |
| DVD | 312 | 28.21% |
| Computer or laptop | 194 | 17.54% |
| Electric iron | 264 | 23.87% |
| Electric blender | 336 | 30.38% |
| Gas stove | 985 | 89.06% |
| Refrigerator | 112 | 10.13% |
| Cloth washing machine | 61 | 5.52% |
| Microwave oven | 41 | 3.71% |
| Sewing machine | 74 | 6.69% |
| Bicycle | 307 | 27.76% |
| Car | 82 | 7.41% |
| Motorcycle | 255 | 23.06% |
| Tricycle | 86 | 7.78% |
| Socio-demographics | The head is employed | 959 | 86.71% |
| The head is a woman | 328 | 29.66% |
| The head is married | 482 | 43.58% |
| The head is literate | 217 | 19.62% |
| The head has no education | 708 | 64.01% |
| The head achieved basic education | 264 | 23.87% |
| The head achieved technic education | 53 | 4.79% |
| The head achieved college education | 51 | 4.61% |
| The head achieved pos-graduate education | 30 | 2.71% |
| The head is a young adult (17 to 35 years) | 103 | 9.31% |
| The head is an adult (36 to 50 years) | 316 | 28.57% |
| The head is an old adult (51 to 65 years) | 361 | 32.64% |
| The head is old (more than 66 years) | 326 | 29.48% |
| Health and insurance  (for household members) | Illness (last month) | 1082 | 97.83% |
| Accident (last month) | 247 | 22.33% |
| Healthy (last month) | 305 | 27.58% |
| Chronic illness | 968 | 87.52% |
| Medical intervention (last month) | 739 | 66.82% |
| Contributory health insurance | 198 | 17.90% |
| Subsidized health insurance | 803 | 72.60% |
| Disabilities | 351 | 31.74% |
| Geographical context | Household is located in a rural area | 670 | 60.58% |

The population of Puno is composed of owned households (82.55%), and they cook using GLP (43.85%) and manure (37.70%). The prevalence of manure cooking is explained by the prevalence of rurality (60.58%), as GLP logistics can be challenging. Is important to notice that only 60.40% of households have daily access to water. Given this context and the high level of exposure to ELEs, it is theoretically logical that population faces high prevalence of respiratory illness, however the categorical features give information at a general level: illness (97.83%), and chronic illness (87.52%). In rural regions over the world, it is common to find that population has health problems (). More than half of the households in sample have at least a member that searched for medical attention (66.82%), and 72.60% of households have subsidized health insurance. The following Figure shows correlations of features that have at least another feature with a correlation higher than 70%. The most correlated features according to this Figure are if ‘rural’, ‘concrete walls’, ‘concrete floor’, ‘drainage network’, ‘water network’, ‘paved tracks’ and ‘paths’. These features can potentially be endogenous and further statistical modelling is needed to draw robust insights about the relationship between these variables and disaster risk. Correlation between features produces multicollinearity, that is addressed by elastic-net regularization for ENLR and by tree-based permutation, that is predictive score or importance in each tree on the ensemble, for RFC. RFC can provide insights about feature importance based on multiple permutations.

Table 6. Continuous features descriptive statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **P50** | **Std** | **Min** | **Max** |
| Household per capita expenditure | 5642.5 | 4307.2 | 4727.5 | 832.4 | 69328.4 |
| Household altitude (m.u.s.l.) | 3836.4 | 3860.0 | 437.2 | 1529.0 | 4835.0 |

Regarding numerical variables, the annual per capita expenditure is measuring short-term household nominal income. The average annual per capita expenditure is S/. 5642.5 nuevos soles from 2017 which is equivalent to 1433$ US dollars at current exchange. The average income is below Latin America principal cities such as Lima, Bogotá, Buenos Aires, Rio de Janeiro. Also, for Puno, the mean income is above the median, which means that more than half of the distribution of per capita expenditure is below the average.

## Hyperparameter optimization

Regarding the hyperparameter optimization approach, the best hyperparameters were selected based on experimental results using a repeated stratified cross-validation scheme. The Table summarizes the results. One characteristic of the proposed solution is that it guarantees a certain robustness of hyperparameters’ configuration as it is based on multiple experiments (k=10) and repetitions (n=2). After NPV optimization, there is no change in ENLR hyperparameters, and for RFC the new parameters are numerically close to the best results of Random Search CV that optimizes MCC alone. For RFC, the change in NPV is greater than the change in MCC. This is important because MCC is the metric that governs classifier performance for binary classification problems (Chicco and Jurman, 2021). A higher AUC suggests that the new model may be more robust to different probability thresholds for prediction. As the data is balanced, the diminution in Accuracy is explained by the reduction in MCC.

Table 7. Hyperparameter configuration before and after NPV optimization

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifier** | **Before NPV optimization** | | | **After NPV optimization** | |
| **ENLR** | **Parameter** | **Value** |  | **Parameter** | **Value** |
| C | 0.100 |  | C | 0.100 |
| l1\_ratio | 0.138 |  | l1\_ratio | 0.138 |
| **Metric** | **Value** |  | **Metric** | **Value** |
| MCC | 54.58 |  | MCC | 54.58 |
| NPV | 80.02 |  | NPV | 80.02 |
| AUC | 82.09 |  | AUC | 82.09 |
| Accuracy | 77.76 |  | Accuracy | 77.76 |
| F1-Score | 81.65 |  | F1-Score | 81.65 |
| **RFC** | **Parameter** | **Value** |  | **Parameter** | **Value** |
| criterion | 'entropy' |  | criterion | 'entropy' |
| max\_depth | 9 |  | max\_depth | 9 |
| max\_features | 0.142 |  | max\_features | 0.141 |
| min\_samples\_leaf | 0.002 |  | min\_samples\_leaf | 0.0001 |
| min\_samples\_split | 0.012 |  | min\_samples\_split | 0.002 |
| n\_estimators | 66 |  | n\_estimators | 87 |
| **Metric** | **Value** |  | **Metric** | **Value** |
| MCC | 56.50 |  | MCC | 56.37 |
| NPV | 81.87 |  | NPV | 82.26 |
| AUC | 82.83 |  | AUC | 82.92 |
| Accuracy | 78.62 |  | Accuracy | 78.53 |
| F1-Score | 82.44 |  | F1-Score | 82.45 |

Metric’s values are computed as the average among folds and repeats within cross-validation scheme.

The Figures 9 and 10 shows the distribution of MCC and NPV metrics for both ENLR and RFC best hyperparameters’ configuration based on the algorithm in Equation 2. Experimental results show a relatively low variability of MCC and NPV across repeats. However, between the folds, there is an important amount of variability. This suggests that the trained model is producing variable results among the data. Considering that the data is a sample drawn from population, this imply that the subset of data that is producing low performance on MCC and NPV could be better modelled by another supervised algorithm. The positive fact is that variability between folds is a pattern, it exists for all the possible configurations of hyperparameters. Future research must seek to minimize the variability between folds, and some algorithms may pay higher attention to mechanisms to minimize this variability.

**Figure 9.** Summary of cross-validation estimates of MCC and NPV for Logistic Regression

|  |  |
| --- | --- |
| **Figure.** Boxplot of MCC over repeats ENLR | **Figure.** Boxplot of NPV over repeats ENLR |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |

**Figure 10.** Summary of cross-validation estimates of MCC and NPV for Random Forest Classifier

|  |  |
| --- | --- |
| **Figure.** Boxplot of MCC over repeats RFC | **Figure.** Boxplot of NPV over repeats RFC |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |

Figures 11 and 12 show the confusion matrix for each algorithm fitted on the best hyperparameters’ configuration that are obtained through the proposed algorithm. The ‘test\_size’ parameter was fixed to 20%, thus model is trained on 80% of sample and tested on the other 20%. For these additional experiments, ENLR achieved a MCC of 63.19% and a NPV of 84.15%. On the other hand, RFC achieved 62.47% and 85.71% respectively.

Nevertheless, these results could be tricky due to the randomness of the split. To overcome this, we repeated the experiments for different values of ‘test\_size’. Results are summarized in Figures 13 and 14. Based on this set of experiments, RFC has a higher change of producing high results with different sizes of train and test subsets. This implies that RFC is producing systematically better predictions than ENLR, and thus is more likely to perform better on real-world applications.

Regarding domain-based hyperparameter optimization (algorithm in Equation 2), the proposed approach led to results with minimum false negatives as is shown in confusion matrices (Figures 11 and 12). Experiments with different test sizes show better results in NPV metrics for RFC (Figures 13 and 14). It is worth mentioning that although differences between the MCC optimization alone and co-optimization of MCC and NPV are small, in real-world applications this would make a difference. When dataset is a sample of a population, assuming that is representative, implementing the model with thousands of inhabitants would lead to important savings in terms of deprivation costs that are important to mitigate risks over the time for recurrent disasters. In this case the sample was designed with the objective of representativity of population, and it also has been previously used to draw insights for policymaking with important impacts (Proaño and Bernabe, 2018).

**Figure 11.** Confusion matrix holdout cross-validation ()

Chart, treemap chart

Description automatically generated

**Figure 12.** Confusion matrix holdout cross-validation ()

Chart, treemap chart

Description automatically generated

Figures 13 and 14 show that, for different test sizes, RFC can adapt better to unseen data as it is producing low-variance performance metrics. Following this line, ENLR has a greater probability of not performing so well as RFC for bigger test sizes. This fact indeed makes a difference in practice, as the cost of misclassifying households at risk of disaster is high because of the potential peaks of deprivation that this could imply.

**Figure 13.** Experiments with different test sizes for Logistic Regression

|  |  |
| --- | --- |
| ENLR |  |
| Chart  Description automatically generated | Chart  Description automatically generated |

**Figure 14.** Experiments with different test sizes for Random Forest Classifier

|  |  |
| --- | --- |
| RFC |  |
| Chart  Description automatically generated | Chart  Description automatically generated |

# Discussion and interpretation

Regarding the results of the training, the proposed strategy for HPO led to good results in terms of performance on test (unseen) data. Furthermore, all the features used for prediction are vulnerability drivers. The main insight of the predictive analysis is that it is plausible to build a good predictive model for disaster risk that is entirely based on vulnerability. This result is important because it states that it is possible to infer where aid is going to be needed whether decision-makers have prior knowledge about geophysical or meteorological characteristics of disasters. It is not true that predictive modeling of cold waves and severe winter conditions based on geophysical-meteorological features is no longer relevant to decision-making, but it is true that proactive measures and interventions to reduce social costs based on open-data empirical modeling would lead to big savings that are important for the disaster risk management cycle, represented as a helix.

The proposed model can be further extended and improved in terms of predictive power, incorporating geophysical-meteorological features such as distance from lakes, rivers, urban settlements. An improvement of predictive power would lead to greater savings, and eventually an optimization of disaster risk management that is focused on proactive PDRRPA. In terms of risk-reduction, we suggest that further statistical analysis and policymaking focus on the most important features that are drawn from model fitting on best hyperparameters configuration. For this case, the features’ importance can be drawn from estimation of RFC on train dataset. The following Figure 15 shows the features’ importance drawn from a single experiment with optimized hyperparameters for RFC:

**Figure 15.** Random Forest Classifier feature importance

Chart, bar chart

Description automatically generated

The insights are clear: most important features for prediction were per capita expenditure (that accounts for short-run household purchase power), household localization in a rural area (that accounts for the fact that household is isolated on the space and systematically far away from principal urban settlements), altitude (that accounts for household exposure to extreme low temperature events), public goods (that can be measuring the presence of the government on public spaces were households are located) and concrete walls (that is capturing the quality of household construction materials). The other features reported on Figure 15 above tell a similar story. Following these results, we confirm a finding that is in line with disaster risk reduction main guidelines: it is necessary to make long-term investment to systematically reduce vulnerabilities to create resilience in communities by achieving socio-economic development of population. Development is a goal that would be achieved at a slow rate, according to historical data there were few examples of rapid development of communities, but these are considered exceptions (cases of study). For instance, human development index tends to evolute slowly over periods of 6 years (Santos et al., 2021). It is worth highlighting the fact that in the short-term, that is the important term for this analysis, machine learning models can be used to minimize resource utilization and, in the best of cases, save important resources that communities may invest in their future development (Bosher et al., 2022).

# Conclusions, recommendations and future research

The main objective of this paper was to discuss the applicability of machine learning based predictive models to solve a humanitarian logistics problem: the proactive supply of aid to a rural community. Additionally, an alternative hyperparameter optimization strategy, to improve solution considering logistics and deprivation costs as multiple objectives, is presented. This strategy is different from state-of-the-art approaches such as Grid Search, Random Search, Genetic Algorithm and other heuristics proposed to find best hyperparameter configuration. The proposed strategy is summarized as follows: optimize by Random Search Cross-Validation considering MCC as the goal in the training process, then from 5% best found hyperparameter configurations pick up the one that produces the highest NPV. MCC metric is important because it accounts for the classification performance considering equal weight for both positive and negative cases. NPV accounts only for negative cases. The main idea behind this is that the 5% best configurations based only on MCC are very likely to produce a result that minimizes the trade-off between misclassification of negative cases and overall misclassification.

The proposed approach gets better predictive performance for negative cases at the cost of a slightly increase in misclassification of positive cases. For humanitarian logistics domain, misclassification of positive cases implies that aid should be delivered to households that are not at risk of being affected by disasters. However, as the majority of Puno’s territory is exposed to cold waves and severe winter conditions it is probably that all the households in population have at least certain degree of risk of being affected by a cold-related disaster, so the delivery of aid to households labeled as ‘non-risk’ could not be unjustifiably increasing costs. This misclassification produces higher logistic costs, but the key assumption behind this analysis is that the reduction in deprivation costs, that comes from accuracy improvement for negative cases, produces more savings than costs caused by the increase in logistic costs, caused by misclassification of positive cases. Thus, the balance of social costs is positive, and this led to important savings considering the case of study that is characterized by a population suffering from high deprivations. For the case of Puno this approach can potentially led to good results, however, the main assumption is only testable by real-world implementation of trained models. For example, in urban areas the savings of the proposed approach may not be as high as in rural case, as urban household are agglomerated in space.

Machine learning offers a solution to the large-scale problem of deciding where aid must be delivered at a disaggregated level. Model predictions can be used to decide what households would require supply of aid. Decision-makers can implement proactive disaster preparedness strategies such as stock pre-positioning (), proactive delivery, and gradual delivery (Apte and Yoho, 2011) based on information drawn from the prediction of trained models. The models can be applied to census data to estimate the magnitude of savings by generating predictions on disaster risks and building an experimental setting. However, model implementation on a context of a real disaster is advisable, considering the objective of measuring savings caused by proactive disaster preparedness strategies applied based on model predictions. The ideal case is to reach an equilibrium between logistic costs and deprivation costs in real-world outcome. Further research will focus on the aspects of model implementation. For future extensions, the recommended pipeline to use SLAs is to train the model with sample data and test the model with real data. The SLAs used in this paper are not scalable to big data, as training time increases logarithmically with number of samples. Testing other SLAs is recommended for future research, for example XGBoost mixes regularization and ensemble, and it is scalable to big data so a big number of experiments can be performed to reach better solutions regarding predictive power of classification metrics. The actual solution achieved an average MCC of 54.58 for ENLR and 56.50 for RFC, and a NPV of 80.02 and 81.87 respectively.

Regarding disaster risk mitigation, this paper confirms the literature findings about vulnerability and disaster risk. Vulnerable households, or deprived households, systematically have a greater probability of being affected by a cold-related disaster. The well-known prescription is to create resilience in communities, which is difficult to achieve in the short term. Instead, we suggest using machine learning to decide where aid must be supplied, considering that humanitarian logisticians operate with scarce resources, and they need to optimize logistics and provide help to communities regardless of their localization or vulnerability condition. Equality on aid distribution can be achieved at a lower cost if aid is delivered in a proactive way, considering that cold-related disasters are seasonal, recurrent and localized in Puno.

References

Santos, R., Santos, P., Sharan, P., & Rodriguez, C. (2021). Digital Agglomeration in the Improvement of the Human Development Index in Peru. In 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC). 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC). IEEE. https://doi.org/10.1109/r10-htc53172.2021.9641710

1. The report states that cold waves happens when temperature drops below 0 °C. [↑](#footnote-ref-1)
2. This represents a conclusion from data analytics presented in Section 2.2 and 2.3. [↑](#footnote-ref-2)
3. Renteria et al. (2021) applied Multiple Correspondence Analysis (MCA) which led their logistic regression model to good predictive power at the cost of smaller reduction ROC-AUC, so in order to increase model adjustment to data, the authors had to apply MCA to a subset of features and re-scale the rest to reach in-sample ROC-AUC of 88% for floods. MCA on high-dimensional data produced information loss that was addressed by applying the dimensionality reduction algorithm to a subset of features. The interpretability of regression results is restricted to a good interpretation of standard coordinates plot (Rencher and Christensen, 2012). [↑](#footnote-ref-3)