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

A predictive assessment of households' risk against disasters caused by cold waves using machine learning ~~techniques and vulnerability predictors~~

Abstract

This paper trains a household-level disaster risk classifier for cold wave-related disasters based on supervised machine learning algorithms. The households' features considered for this task ~~come from~~proxies~~proxies~~proxy ~~multidimensional~~ multiple dimensions of vulnerability to disasters accounting for economic, health, social, and geographical dimensions. These features are theoretically hypothesized to explain disaster risk classification. We test our predictive model based on the case of Puno, Peru, where cold wave-related disasters (e.g., ~~2003~~ -28°in 2003 and ~~, 2004~~ -35° in 2004) are recurrent and overwhelming. To build the classifier, two supervised learning algorithms were tested: Logistic Regression and Random Forest Classifier. Hyperparameters of such models were optimized through a heuristic applied to results of Random Search Cross-Validation, such as the configuration maximize the 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. In the optimal setting, Negative Predictive Value, which controls the false negatives, increased without trading off a significant amount of model performance as suggested by MCC and other metrics. ~~Based on~~Considering experiments with different sizes of test datasets, the optimal Random Forest Classifier outperformed the optimal Logistic Regression classifier. Feature importance drawn from features' contribution to a reduction in entropy in the construction of the forest suggests that per capita expenditure, household localization in a rural area, altitude, access to public goods, and concrete walls drive the disaster risk classification. Further research must propose strategies to validate the predictive model externally and to analyze the causality of the most important features regarding endogenous disaster risk classification.

1. 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, the scale and magnitude of the hazard, and the level of exposure. Hazards 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 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).

Commented [AL1]: tente pensar em um título mais simples

Commented [AL2]: Sugiro retirar essa parte do título

Commented [AL3]: cold-related? É assim que a literatura se refere? Para mim não é claro sobre o que você fala aqui. Seria climate-related?

Commented [AL4R3]: cold wave-related?

Commented [RJQA5R3]: A base de dados registra como cold waves ou severe winter conditions,

A literatura fala de cold waves

Vou ficar com cold waves

Commented [AL6]: Adicionar título e and abstract

Commented [AL7]: Tente reestruturar esta seção melhorando o storyline. Cada parágrafo tem um o intuito de apresentar uma ideia específica. E todo parágrafo está conectado com o anterior.

Tente reformular o texto indo do mais geral para o mais específico.

A estrutura deve ser:
motivação
objetivos geral e específico
visão geral de como atingiu o objetivo (metodologia)
contribuição
estrutura das demais seções

Commented [FA8]: Eu comecaria a introducao com este paragrafo, que apresenta o contexto do trabalho, antes de mencionar o repositório de dados e detalhes mais específicos

Commented [FA9]: Falta referencia

Proactive disaster risk reduction is essential for communities affected by recurrent disasters. However, disaster risk management phases are not independent of each other (Besiou et al., 2021). Thus, proactive disaster risk reduction activities that are carried out before a disaster strike 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.

The increase in the frequency of cold wave-related disasters during the last century disproportionately affected low-income countries (Amirkhani et al., 2022; Lopez-Bueno et al., 2020). India, Bangladesh, Poland, and Russia are the most-affected countries , harming 1227 million people and generating 184 thousand deaths since 2000. Disasters triggered by cold waves cause losses of human lives in cases of high vulnerability, where households have poor infrastructure and scarce goods to face cold (Lopez-Bueno et al., 2020) or inhabitants have a high prevalence of comorbidities such as cardiovascular diseases (Shaposhnikov and Revich, 2016).

This work focuses on disaster preparedness following a data-centric approach (EM-DAT, 2022). We aim 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, which represent demand points that must be attended. When a predictive model misclassifies positive outcomes, defined as households at risk, deprivation costs are created to represent demand points that need essential supplies, but the model misclassifies their risks, and aid goods are not being supplied (Gutjahr and Fischer, 2018; Holguin-Veras et al., 2013). These cases are named false negatives.

Commented [AL12]: What risks cold waves bring to households? Need to explain it!

Our proposed model gives 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, our methodology uses supervised learning algorithms - Logistic Regression and Random Forest Classifier - with data from the 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 wave-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.

Puno, in Peru, is affected by recurrent cold waves. 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 (2022), 21.28% of them have affected this geographic boundary. According to EM-DAT estimations, the most harmful ELTE was recorded in 2004 as a cold wave of -35°C that affected 40.30% of the total population of 15 Peruvian regions. Puno is a rural and low-densely populated region located in the southeast of Peru. Puno is the epicenter of ELTEs affecting PSAR, as 70.00% of events registered in EM-DAT affected Puno from 2003 to 2015. As ELTEs affect a large geographic boundary, it could be challenging to estimate the number of affected people, the economic losses, etc.

Commented [AL13]: Não crie tantas siglas no texto. É difícil entender.

Research on proactive disaster risk reduction would greatly impact Puno because of the high Research on proactive disaster risk reduction would greatly impact Puno because of the high prevalence of agricultural households', in which these disasters may cause economic losses that impact their long-run wealth. If a community is not prepared to face cold wave-related disasters, it might enter

into a vicious cycle of cold waves affecting the economy. This vicious cycle affects the ability to respond and recover from disasters, producing a lower budget to invest in resilience mechanisms (Besiou et al., 2021).

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 include 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); thus, 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 the main concern of humanitarian logistics.

The contribution of this paper is twofold. First, we introduce vulnerability-based disaster risk prediction, in contrast with other predictive strategies based on meteorological, geophysical, or geographical modeling. Then, we propose a hyperparameter optimization algorithm based on domain requirements, such as minimizing false negatives. The key element for the hyperparameter optimization procedure is the confusion matrix of the predictive models, as logistics costs depend on False Positives and True Positives. True Negatives mean no delivery is required, and deprivation costs arise from False Negatives. The experimental setting for hyperparameters' optimization considers confusion matrix metrics by co-optimizing on Matthews Correlation Coefficient (MCC) and Negative Predictive Value (NPV). HPO is usually based on one metric, but our methodology 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 reducing 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 on where and at which level to preposition supplies is high, as those supplies aim to reduce the expected damages to households' livelihoods that are strongly linked to agriculture and livestock (Quiliche and Mancilla, 2021).

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 emergency assessment. Section 3 details the data collection methods and processing pipeline and experimental setting. Section 4 brings the results for hyperparameter optimization and deprivation costs. Section 5 discusses the main results. Finally, Section 6 brings our conclusions and recommendations for improvements in disaster preparedness strategies and future research avenues.

Theoretical foundation

In this section, we first discuss disaster risk reduction (DRR) concepts. Then, we overview the data science applications in DRR.

Commented [LA14]: não aqui. Coloque isto na conclusão, como proposta de trabalhos futuros.

Commented [AL15R14]: Faltou levar isto para a conclusão.

Commented [FA16]: Temos que ter cuidado com as afirmações muito fortes, sugiro suavizar o discurso (ao longo de todo o texto)

Commented [LA17]: Esse tem que ser o foco da seção 2. Não focar em Puno nesta seção

2.2.2.1. Disaster risk reduction for clime-related disasters

The most outstanding theory on disaster risk claims that risk is produced if three elements are combined for a geographic boundary (Mors, 2010; UNDRR, 2015; Twigg, 2004): i. natural hazard, i.e., the natural phenomenon that may harm communities; ii. exposure, i.e., the condition of an agent within the geographic boundary of being exposed to such natural hazard; and iii. vulnerability, which shapes the consequences of a damaging event on agents. If an agent is resilient to disasters, then it would have small losses after a disastrous event. Vulnerability is a set of conditions that an agent posse that make it more prone to high losses when it is affected by a hazardous event (Christian et al., 2021; Sahana et al., 2019; Tasnuva et al., 2020; Ullah et al., 2021). Among natural hazards that jeopardize vulnerable communities, clime-related hazards such as rainfalls, heat waves, cold waves or storms have an impact that covariate with the degree of vulnerability of the agents within the geographic boundary exposed to such hazards (Renteria et al., 2021).

However, this premise is not directly generalizable for all types of disasters. For example, earthquakes imply greater uncertainty regarding losses, and this means a different relationship between disaster risk and vulnerability. Building resilience for earthquakes may require additional efforts that are beyond the scope of this research. In contrast, the adverse effects of recurrent clime-related hazards can be mitigated as the requirements in terms of risk reduction are simpler. Furthermore, these hazards tend to be seasonal, localized in a geographic boundary and the magnitude of losses can be relatively easy to anticipate (Simmons and Sutter, 2014).

The importance of disaster risk reduction comes from vulnerability shaping the magnitude of the losses related to agents' exposure to natural hazards. Disaster risk can be mitigated by the reduction of vulnerability, or equivalent, by the creation of resilience, as stated in the Sendai Framework for Disaster Risk Reduction (Aitsi-Selmi et al., 2015).

Reduction of vulnerability is a long-term goal. From an economic perspective, communities need resources to face disasters. Furthermore, when a community is affected by recurrent disasters, disaster risk reduction could be especially challenging. In those cases, the resources allocated to response and recovery from disasters are more likely to be higher than resources invested in risk mitigation and disaster preparedness. Thus, the total cost of the disaster risk management cycle is steadily high, as illustrated by the red line in Figure 1.

In this regard, Bosher et al. (2021) states that pre-disaster risk reduction and preparedness activities must aim to reduce the total cost of the disaster risk management lifecycle. If a community successfully builds resilience through proactive interventions on disaster risk reduction and preparedness, then future disasters would produce lower losses. In such cases, the total cost can be smoothed as it is illustrated by the green line in Figure 1.

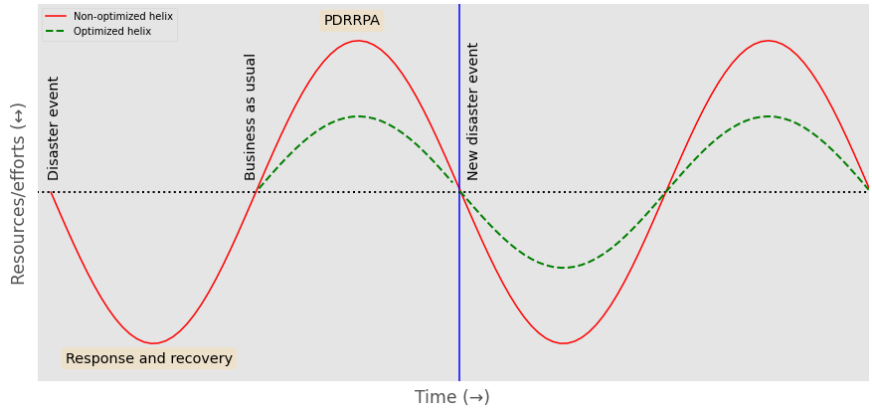
Figure 1. Theorical representation Disaster Management Helix optimization

Commented [AL18]: Esta seção é sobre aplicação de data science em DRR. Estes parágrafos, se necessários, deveriam estar na seção anterior.

Commented [AL19R18]: A ideia da seção é mostrar quais aplicações de data science para DRR existem e o que você traz de contribuição. Acho que você foca muito em descrever o que estamos propondo. Isto pode ser feito em um parágrafo ao final da seção, mas não é o objetivo da seção.
Precisa trabalhar um pouco mais neste texto.

Commented [AL20]: Dá onde veio essa figura? Não copiar figuras de outras fontes. É plágio.

Commented [AL21R20]: retirar a figura ou adaptá-la



Adapted from Bosher et al. (2021)

The concept of disaster risk management helix illustrates better the dynamics of disaster risk reduction. The long-term matters when a community faces recurrent disasters. In cases where communities are affected by recurrent disasters, disaster risk might harm the overall economic environment by having infrastructure destruction, systemic agricultural losses, and hazard to public health (Ferreira, 2012; López-Bueno et al., 2021; Quiliche and Mancilla, 2021).

The first contribution of this paper is that it proposes a solution that is aware of the long-term dynamics of the disaster risk management lifecycle. The implementation of a Machine Learning classifier of this type aims to anticipate disaster-related losses in order to target policies to mitigate risks and prepare agents for upcoming disasters.

2.2. Machine learning in disaster risk reduction

Previous studies addressed disaster preparedness with predictive analytics (Davis et al., 2010; Simmons and Sutter, 2014; Van Thang et al., 2022). There are several contributions of Machine Learning to disaster risk management. Lu et al. (2021) performed a comprehensive review of applied Machine Learning in context of public health emergencies related to disasters. The authors found that the main contribution of Machine Learning is to process information to support decision-making in management of risks by producing forecasts and insights to improve understanding of phenomena. For example, automated models can improve decision-making under time-sensitive conditions by processing big data. In this sense, Machine Learning contributes to multiple edges of information management: demand forecasts may help to reduce material convergence (Holguin-Veras et al., 2014), stochastic programming in transportation may help to avoid bottlenecks (Alcántara-Ayala, 2019), and so on. Machine learning not only helps to predict, but it also helps to understand complex phenomena. For instance, data mining applied to disaster risk management is defined as the process where algorithms find insightful patterns in data that represent chaotic environments characterized by high uncertainties (Fayyad and Shapiro, 1996; Tomasini and Van Wassenhove, 2009; Behl and Dutta, 2018). In Perú, Izquierdo-Horna et al. (2022) applied a hybrid approach to seismic risk assessment, integrating Random Forest and Hierarchical Analysis to determine seismic risk in Pisco. China is a country known for having densely populated cities. An early-awareness

Commented [AL22]: Esta seção é sobre aplicação de data science em DRR. Estes parágrafos, se necessários, deveriam estar na seção anterior.

Commented [AL23R22]: A ideia da seção é mostrar quais aplicações de data science para DRR existem e o que você traz de contribuição. Acho que você foca muito em descrever o que estamos propondo. Isto pode ser feito em um parágrafo ao final da seção, mas não é o objetivo da seção.
Precisa trabalhar um pouco mais neste texto.

approach based on machine learning is beneficial in that context. The disaster response plan can be executed within a more extended time window before flooding is at its peak (Bai et al., 2022).

A critical gap identified in machine learning applications is that predictive modeling, in the cases where it incorporates vulnerability, characterizes vulnerability by economic factors. A multidimensional approach needs to be included to represent vulnerability better. This multidimensional vulnerability approach contributes to a better understanding of clime-related disaster risks and improves prediction accuracies (Mors, 2010).

Regarding disaster risk understanding, few studies have considered comprehensive data on multidimensional vulnerability (for example, Ahmad and Routray, 2018; Patri et al., 2022.). The dimensions of vulnerability are composed of variables with endogenous nature. These features are likely to covariate with other predictors not considered in this paper. For example, vulnerable agents tend to be settled in places with high exposure. The classifier is expected to exploit these relationships to produce accurate predictions. In short, the amount of information that multidimensional vulnerability features provide makes it feasible to train an accurate classifier from the vulnerability characterization of agents.

To summarize some previous findings: low-income and lousy infrastructure are the main drivers of vulnerability to clime-related 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).

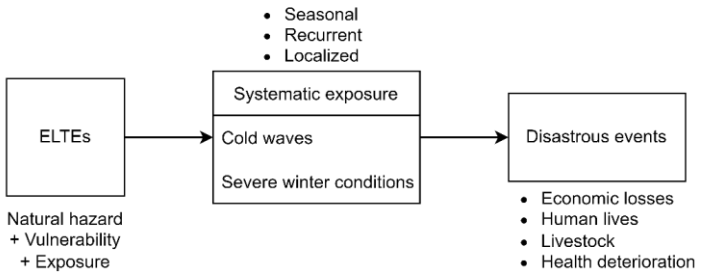
Our approach is grounded on previous results summarized in exploratory statistical analysis for clime-related disasters (López-Bueno et al., 2021; Renteria et al., 2021). The probability that an agent would be affected by a natural hazard increase when a set of characteristics are met, such as lack of access to basic services, lack of health, low educational achievement, low social development (Pessoa, 2012) and geographical exposure for the case of disasters (Ullah et al., 2021). In this paper, vulnerability has four dimensions: economic, health, social and geographical. The second contribution of this papers is that it explores an expanded feature space for vulnerability that considers multiple dimensions that may explain disaster risk.

3. The case of Puno, Perú

Commented [AL26]: essa seção não é a 2.3, mas a 3

The third contribution of this paper is that it adapts the standard Machine Learning pipeline to a particular case of study: the Puno region of Peru. The Puno region is affected by cold wave-related disasters. Cold waves reach large geographic boundaries. In such cases, where the entire population is exposed to hazardous events, differences in vulnerability shapes differences in disaster risk. Disasters are more likely to happen where households are more vulnerable. In this case, cold waves produce higher losses for agricultural households, or households that are built with low quality materials. Figure 2 illustrates the triggering process of cold waves-related disasters.

Figure 2. Causes of cold-related disastrous events affecting households in Puno



Formatted: Centered

Regarding cold-wave related mortality, López-Bueno et al. (2021) performed statistical analysis of mortality rates 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 have 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.

Table 2 shows an event log of cold wave-related disasters that affected the south Andean region of Perú including Puno. These logs are representative of the magnitude of the cold waves in terms of minimum temperature, duration, affected people and year when the disaster was registered. This information along with the time series of minimum temperature reported in Figure 3 provides evidence that illustrates the seasonality of the cold waves in Puno. Every year, households located within Puno are exposed to cold waves. In July, August and September, the exposure tends to be higher on average for all the meteorological stations that collect temperature data in Puno.

Table 2. Event log of cold wave-related disasters that affected Puno

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.

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

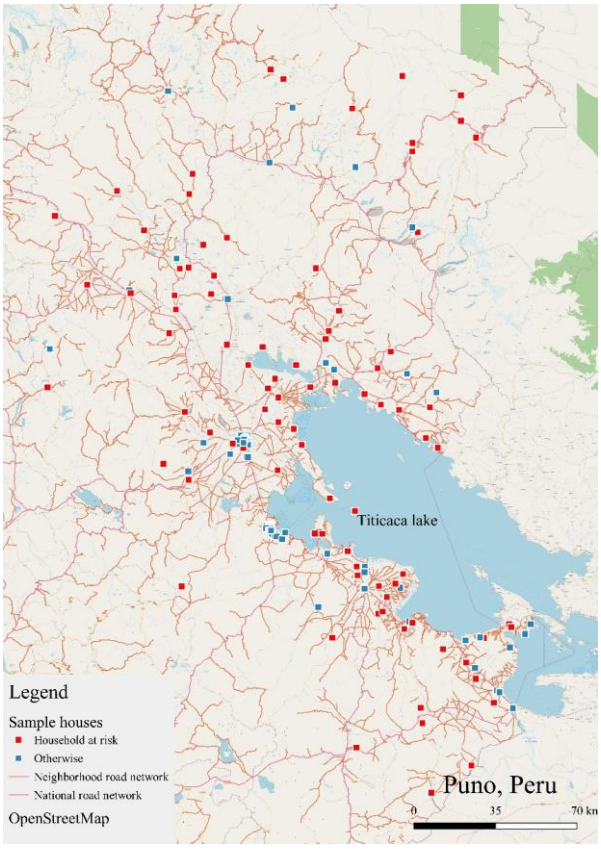
an average of 15 days for June, July and August within Puno. In this report, the authors conclude that during the period 1969-2010, for each year there was at least one cold wave[1], this does not mean that for each instance a disastrous event was triggered. The historical data about disastrous events 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).

The analysis was made at the household level. This level of granularity allows the researchers to draw insights for the points of final demand of aid, also called the final echelon of the Humanitarian Supply Chain (Chong et al., 2019). Such information is valuable for the development of humanitarian operations including disaster preparedness strategies.

A particular characteristic of the case of study, that matters because the goal of the classifier is to identify the final demand points, is that population are settled dispersed in space. In Figure 4, the localization of the majority of final demand points is outside the principal cities, in rural areas. In those cases, the logistic costs are high (Gutjahr and Fischer, 2018). On the other hand, misclassifying households that are at risk of being affected by a cold wave-related disaster would produce deprivation costs because those households do need aid, but the model discriminates that they do not (Tomasini and Van Wassenhove, 2009; Leiras et al., 2017). Leaving the problem of high logistic costs for future research, as transport is outside the scope of this research, a greater penalization for false negatives was incorporated at the model training stage in order to produce accurate classifications with minimum deprivation costs.

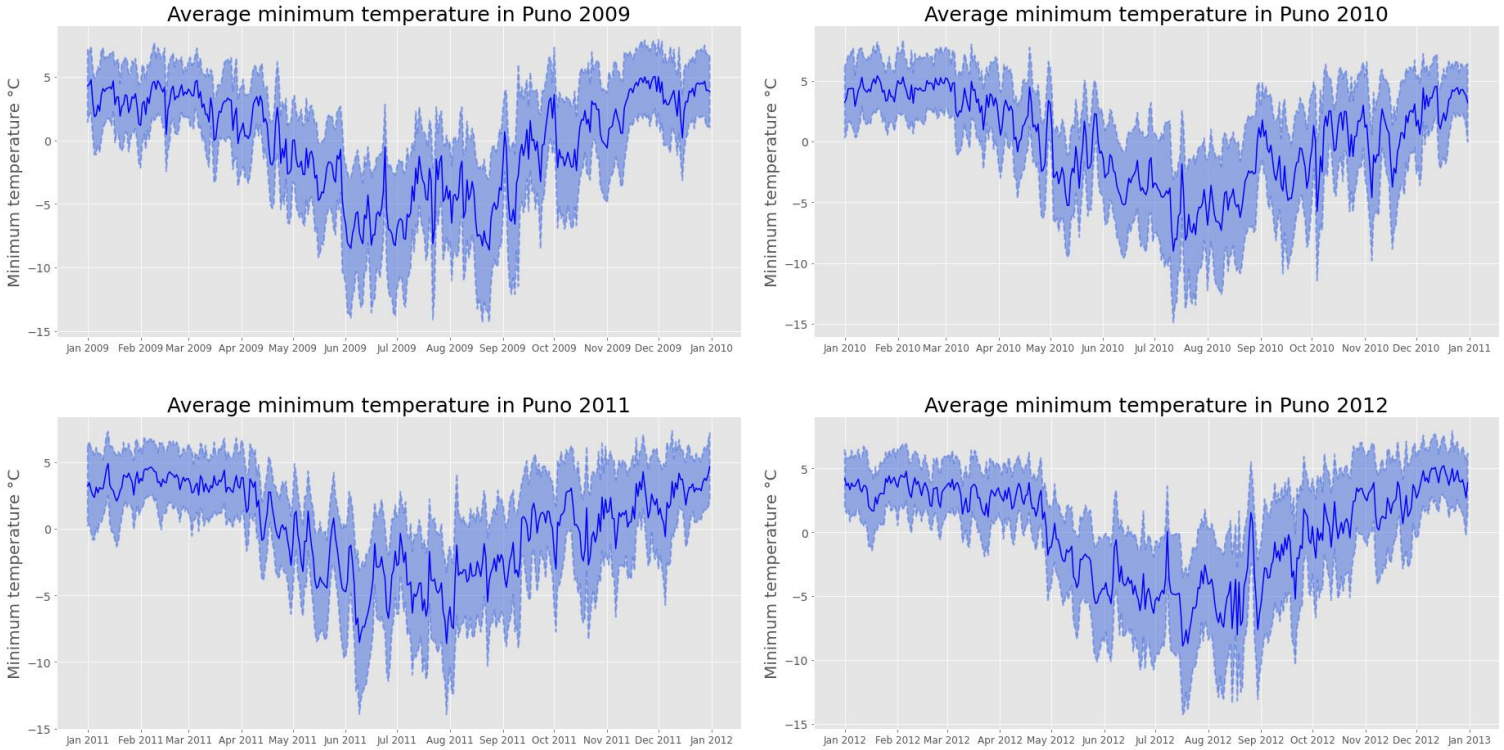
Figure 4. Spatial distribution of households exposed to ELTEs

[1] The report states that cold wave-related disasters happens when temperature drops below 0 °C.



This paper part from the problematic of disaster risk reduction for communities with recurrent disasters. Then it proposes to train a classifier using Machine Learning methods in order to identify points of final demand and support humanitarian operations. The approach focusses on supporting pre-disaster risk reduction and planning activities, as it is the phase where the greater impact of model implementation is expected. Nevertheless, the insights may be useful for post-disaster response and recovery activities, as it also contributes to the understanding of vulnerability drivers at the household level.

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



~~1.~~4. Materials and methods

Formatted: Indent: Left: 0", Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

1.1. Data collection methods and the classification problem

Raw data on households' dimension of vulnerability to disasters were collected from the National Household Survey (NHS) that was carried out by Peruvian's National Institute of Statistics and Informatics in 2019. Data is available at the nationwide level. The sampling method was stratified over political regions. In consequence, the survey is representative of Puno at regional level. The following survey modules were considered for this analysis: population and housing (modules 100 and 200), education (module 300), health (module 400), employment (module 500), and democracy and transparency (module 612). These modules contain information about the defined dimensions of vulnerability (UNDRR, 2015; Salazar-Briones et al., 2020 and Renteria et al., 2021). In the survey, 23.33% of questions were answered by another informant that is not the head of the household.

The learning target is a binary indicator, as shown by Equation 1. The empirical classification problem will be assessed through supervised learning techniques. Regarding classification classes, module 612 asks the following question for each household:

In the last 12 months, has your house been affected by natural disasters (drought, storm, plague, flood, etc.)?

This question does not provide specific information about the type of disaster associated with cold wave-related risk. The following facts must be considered to argue for the appropriateness of this variable to measure households' risk of being affected by a cold waves-related disaster:

1. In the specific case of Puno, there is an overwhelming prevalence of risks related to low temperatures (see Section 2.3 for data analytics support for this proposition).
2. The average household's monthly earnings are S/. 470.2, and the poverty line is estimated on S/. 352. Based on data from the National Household Survey, 48.6% of households were poor in 2019 and, thus, very likely to be affected by cold waves-related disasters.
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 clime-related disasters. For example, if a household is at risk of being affected by floods, it is very likely to be 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 cold waves-related disasters. Supervised learning will use this variable as the target for the classification problem.

$$Y_i = \begin{cases} 1 \text{ if household is at risk of being affected by a cold-related disaster} \\ 0 \text{ otherwise} \end{cases}$$

Equation 1. Binary variable measuring household disaster risk

Commented [FA27]: Esta diferente da introducao

1.2. Machine learning pipeline

All the Machine Learning pipeline steps were performed on Python 3.10 programming language using Scikit-Learn 1.1.1 package.

Data pre-processing

The feature space extracted from NHS is multi-dimensional. This strategy overcomes the empirical over-simplification of disaster vulnerability into the socio-economic dimension and ignores the other factors' dependence (Villarroel-Lamb, 2020; Regal, 2021; Szczyrba et al., 2021 are some examples). However, the proposed approach entails greater empirical complexity as a higher number of features are considered for the model training process:

"High-dimensional datasets bring a lot of information to people, but 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).

The dataset comprises 86 features, 84 are binary, and the rest are numeric. A higher number of features will increase the computational time required for hyperparameter search for each supervised learning algorithm. Standard dimensionality reduction techniques, such as Principal Component Analysis, resulted in a significant loss of information that affected the predictive power of supervised algorithms in preliminary experiments. To overcome the curse of dimensionality, supervised sparse learning algorithms were selected: Random Forest and Elastic-Net Logistic Regression. These algorithms rank the feature's importance and reach the optimal predictive formula as a function of a subset of features, removing large amounts of redundancy and noise in the dataset (Xuan et al., 2019).

Following packages' documentation guidelines, selected supervised learning algorithms' performances improve when input features are measured on the same scale. Consequently, features were standardized by removing the median and scaling the data according to the interquartile range. This feature standardization process is known as Robust Scaling. This procedure was selected because the distribution of numeric variables is asymmetric and does not fit a standard normal distribution. For instance, per capita expenditure is known to be right-skewed, as purchase power tends to have high inequalities in developing countries.

Data processing

Instead of generating information loss, which is the case for standard dimensionality reduction methods, sparse learning includes features' regularization terms into their objective functions. It reaches maximum predictive power by balancing the bias-variance trade-off through a cross-validation process that aims to maximize a performance metric (Jian et al., 2008; Robert, 2011). Regularized logistic regression includes an Elastic-Net regularization, a convex combination of L1 and L2 regularization. On the other hand, Random Forest is a non-parametric technique that ranks features according to their contribution to the reduction in Gini or Entropy of the conditional probability distribution of the outcome. If a feature does not contribute to an improvement in classification performance or if it does not produce information gain, then that feature is discarded as a candidate for a split (Xin and Ren, 2022).

For the hyperparameter optimization strategy, the repeated K-fold cross-validation method

was done using the Random Search Cross-Validation method.

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:
   3.1. 'Min_samples_split': minimum number of data points placed in a node before the node is split.
   3.2. '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:

$$\min_{\beta,\beta_0} \frac{1-\rho}{2}\beta^T\beta + \rho\|\beta\| + C\sum_{i}^{N}\log\left(\exp\left(-Y_i\left(x_i^T\beta + c\right)\right) + 1\right)$$

Where x_i^T is a data vector corresponding to observation i, Y_i 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:

$$F\left(x_i^T\right) = sgn\left(\sum_{k=1}^{K} Y_i\,\beta_k K\left(x_i, x_i^T\right) + \beta_0\right)$$

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

Commented [FA29]: ?

to regularize parameters to reach sparse results on coefficients β_k. 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:

$$K(\mathrm{x}_i, \mathrm{x}_\mathrm{i}^\mathrm{T}) = \exp\left(-\gamma\left|\mathrm{x}_\mathrm{i} - \mathrm{x}_\mathrm{i}^\mathrm{T}\right|^2\right)$$

1.1. 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

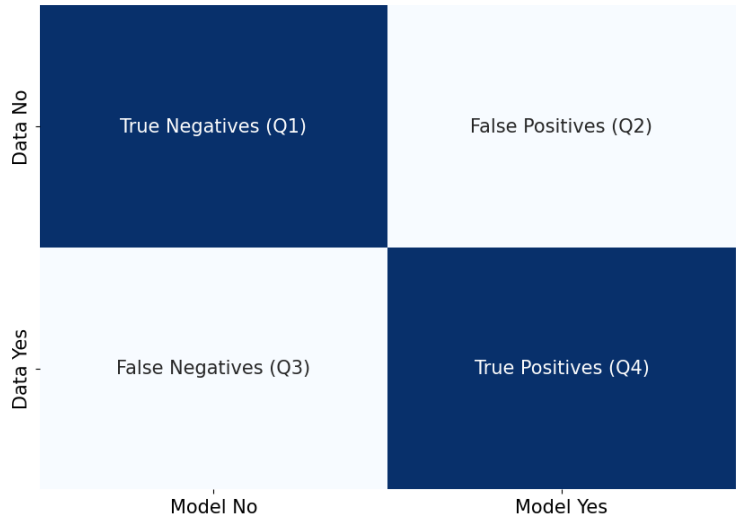


Table 3: Interpretation of confusion matrix elements

Quadrant	Relation with social costs
$TN\ (Q1)$	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.
$FP\ (Q2)$	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 $(O2)$.
$FN\ (Q3)$	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.
$TP\ (Q4)$	Cases where households have risk and the model decides that they must be delivered with aid, thus generating justified logistic costs ($O2$)

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 $TruePositiveRate = \frac{TP}{TP+FP}$ and $FalsePositiveRate = \frac{FP}{TP+FP}$ coordinates at various probability threshold settings. The range of this metric varies in the closed interval $[0,1]$ so better classifiers are found when $AUC \to 1$.

Accuracy

The estimation of accuracy represents the application of common heuristic where diagonal of confusion-matrix is maximized. The formula is given by $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$. The range of this metric varies in the closed interval $[0,1]$ so better classifiers are found when $Accuracy \to 1$.

F1-Score

Is defined as the harmonic mean of the $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$. The formula is given by $F1 = \frac{TP}{TP+0.5(FP+FN)}$. The range of this metric varies in the closed interval $[0,1]$ so better classifiers are found when $F1 \to 1$.

Matthews Correlation Coefficient

This metric is in essence a correlation coefficient that lays in the [-1,1] interval. The formula is given by $MCC = \frac{TP(TN)-FP(FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FP)}}$. 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 $NPV = \frac{TN}{TN+FN}$. The proposed HPO strategy will try to co-optimize MCC and NPV.

1.2. 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

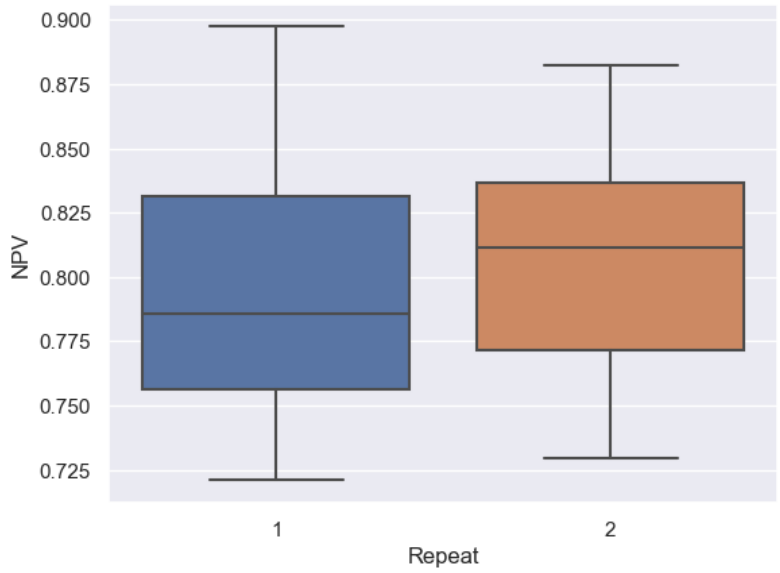
Commented [FA30]: Poderia colocar na notacao de algoritmo mesmo, com chamada, parametros, etc

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	criterion=['gini', 'entropy'] max_depth=randint(2,10) max_samples=['0.2', '0.5', '0.8'] min_samples_split=uniform[0, 0.6] min_samples_leaf=uniform[0, 0.5] n_estimators=randint[10,300] max_features=['log2', 'sqrt']
Elastic-Net Logistic Regression	C=[0.01, 0.1, 1, 10, 100] l1_ratio=uniform[0,1]
Extreme Gradient Boosting (XGBoost)	
Support Vector Classifier (SVC)	C=[]



2.5. 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.

2.1.1.3. Descriptive statistics

Figure 8: Features' correlation heatmap

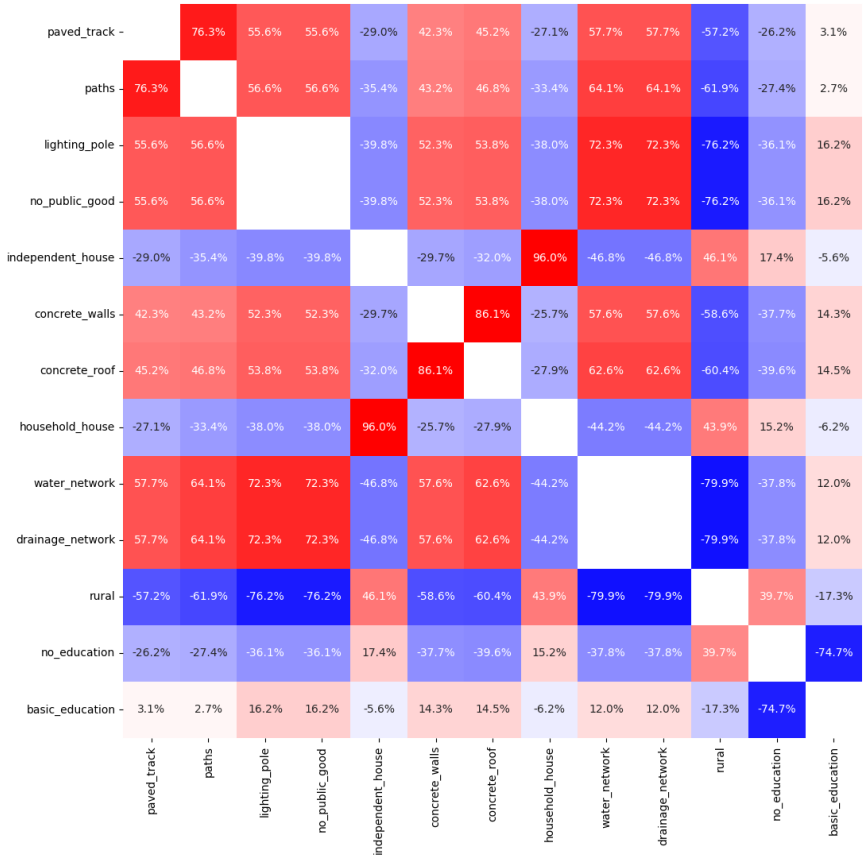


Table 5. Descriptive statistics

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.

~~2.2.~~1.4. 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

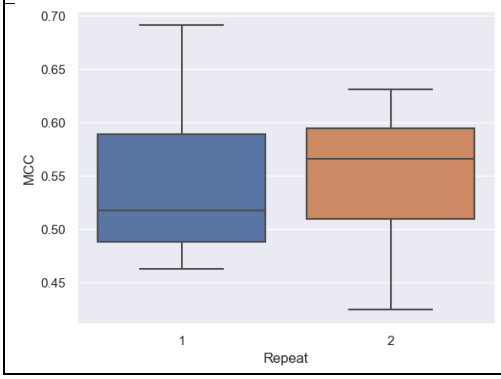


Figure. Boxplot of NPV over repeats ENLR

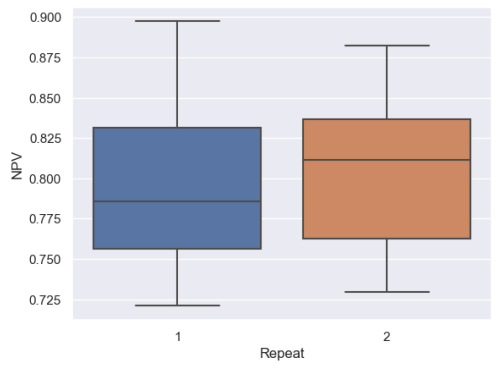


Figure 10. Summary of cross-validation estimates of MCC and NPV for Random Forest Classifier

Figure. Boxplot of MCC over repeats RFC

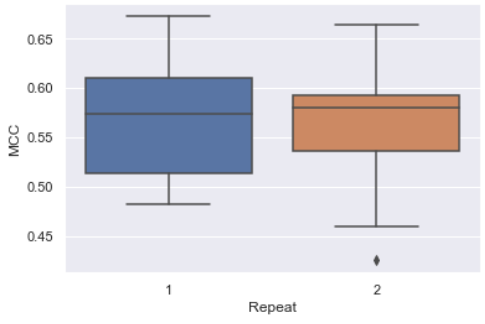
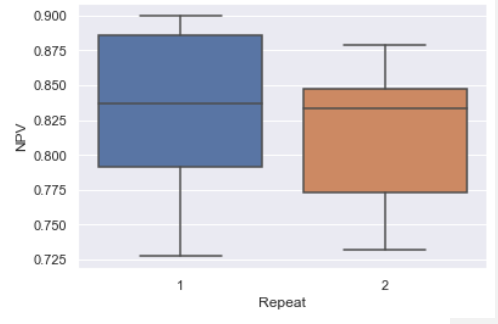


Figure. Boxplot of NPV over repeats RFC



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.

Commented [FA31]: ?

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 ($test\_size = 20\%$)

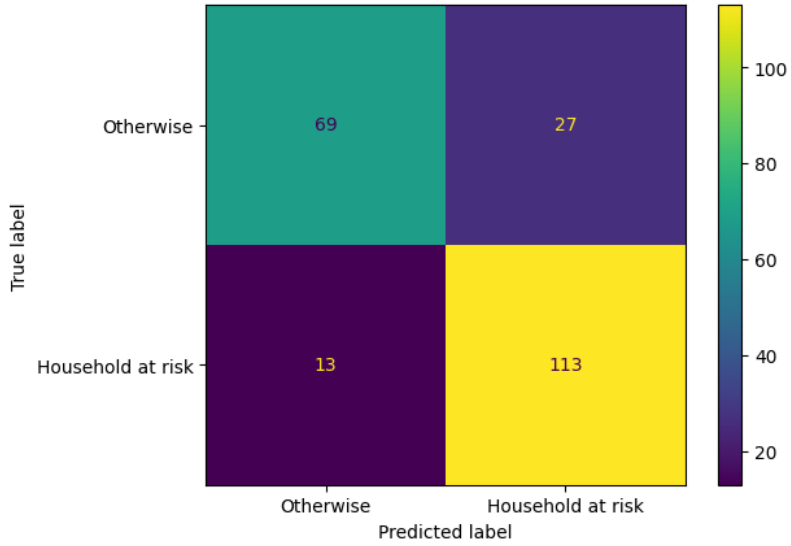
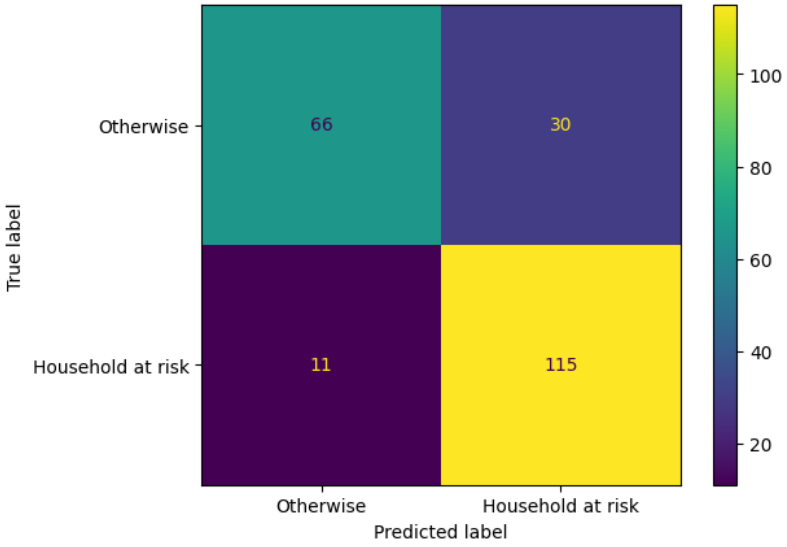


Figure 12. Confusion matrix holdout cross-validation ($test\_size = 20\%$)



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.

Commented [FA32]: ?

Figure 13. Experiments with different test sizes for Logistic Regression

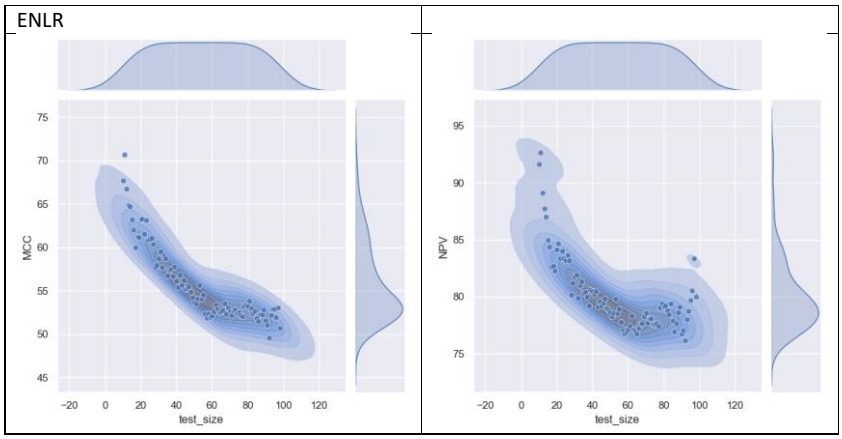
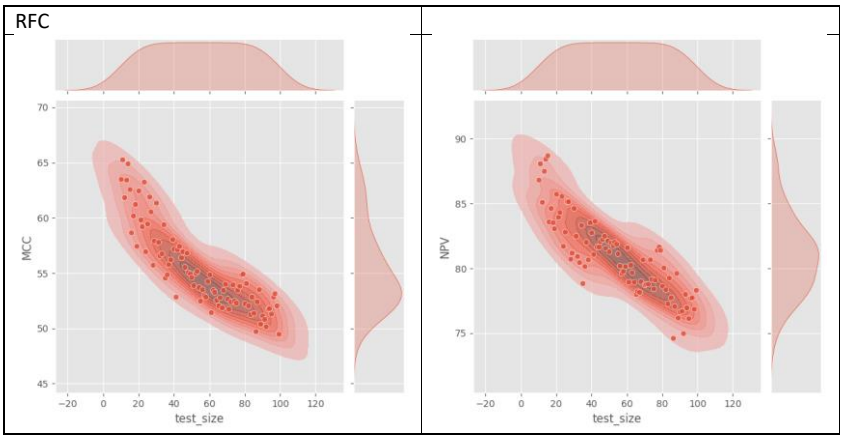


Figure 14. Experiments with different test sizes for Random Forest Classifier



~~3.~~6. Discussion and interpretation

Formatted: Indent: Left: 0", Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

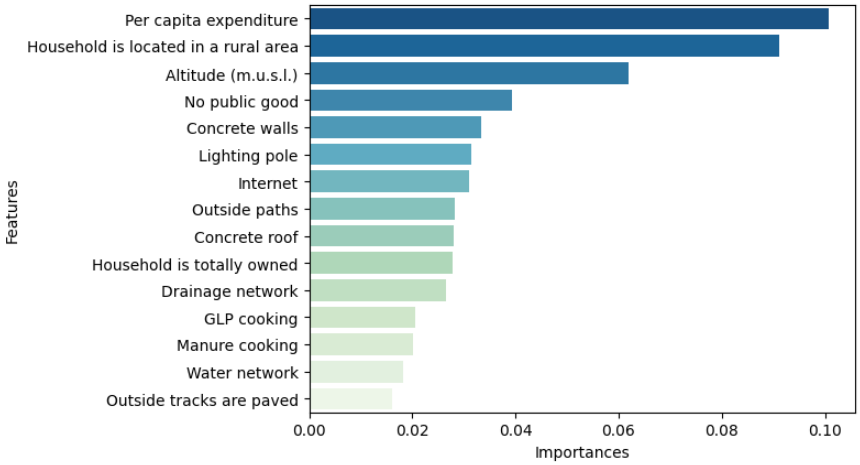
Regarding the results of the training, the proposed strategy for HPO led to good results in terms of performance on test (unseen) data. Furthermore, all the features used for prediction are vulnerability drivers. The main insight of the predictive analysis is that it is plausible to build a good predictive 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

Commented [FA33]: Numero?



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).

4.7. 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

Ahmed Arafa, A., Radad, M., Badawy, M., & El-Fishawy, N. (2022). Logistic Regression Hyperparameter Optimization for Cancer Classification. *Menoufia Journal of Electronic Engineering Research*, *0*(0), 0–0. https://doi.org/10.21608/mjeer.2021.70512.1034

Amirkhani, M., Ghaemimood, S., von Schreeb, J., El-Khatib, Z., & Yaya, S. (2022). Extreme weather events and death based on temperature and CO2 emission – A global retrospective study in 77 low-, middle- and high-income countries from 1999 to 2018. *Preventive Medicine Reports*, *28*, 101846. https://doi.org/10.1016/j.pmedr.2022.101846

Balcik, B., Beamon, B. M., Krejci, C. C., Muramatsu, K. M., & Ramirez, M. (2010). Coordination in humanitarian relief chains: Practices, challenges and opportunities. *International Journal of Production Economics*, *126*(1), 22–34. https://doi.org/10.1016/j.ijpe.2009.09.008

Bosher, L., Chmutina, K., & van Niekerk, D. (2021). Stop going around in circles: towards a reconceptualisation of disaster risk management phases. *Disaster Prevention and Management: An International Journal*, *30*(4/5), 525–537. https://doi.org/10.1108/DPM-03-2021-0071

Cantillo, V., Serrano, I., Macea, L. F., & Holguín-Veras, J. (2018). Discrete choice approach for assessing deprivation cost in humanitarian relief operations. *Socio-Economic Planning Sciences*, *63*, 33–46. https://doi.org/10.1016/j.seps.2017.06.004

Daumé, H., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, *26*, 101–126. https://doi.org/10.1613/jair.1872

Gutjahr, W. J., & Fischer, S. (2018). Equity and deprivation costs in humanitarian logistics. *European Journal of Operational Research*, *270*(1), 185–197. https://doi.org/10.1016/j.ejor.2018.03.019

Hasebrook, N., Morsbach, F., Kannengießer, N., Franke, J., Hutter, F., & Sunyaev, A. (2022). *Why Do Machine Learning Practitioners Still Use Manual Tuning? A Qualitative Study*. http://arxiv.org/abs/2203.01717

Holguín-Veras, J., Jaller, M., van Wassenhove, L. N., Pérez, N., & Wachtendorf, T. (2014). Material Convergence: Important and Understudied Disaster Phenomenon. *Natural Hazards Review*, *15*(1), 1–12. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000113

Holguín-Veras, J., Pérez, N., Jaller, M., van Wassenhove, L. N., & Aros-Vera, F. (2013). On the appropriate objective function for post-disaster humanitarian logistics models. *Journal of Operations Management*, *31*(5), 262–280. https://doi.org/10.1016/j.jom.2013.06.002

Huang, X., Wu, L., & Ye, Y. (2019). A Review on Dimensionality Reduction Techniques. *International Journal of Pattern Recognition and Artificial Intelligence*, *33*(10), 1950017. https://doi.org/10.1142/S0218001419500174

Kim, Y., & Sohn, H.-G. (2018a). *Disaster Risk Management in the Republic of Korea*. Springer Singapore. https://doi.org/10.1007/978-981-10-4789-3

Kim, Y., & Sohn, H.-G. (2018b). *Disaster Risk Management in the Republic of Korea*. Springer Singapore. https://doi.org/10.1007/978-981-10-4789-3

Li, X., Caragea, C., Caragea, D., Imran, M., & Ofli, F. (2019). Identifying disaster damage images using a domain adaptation approach. *Proceedings of the International ISCRAM Conference*, *2019-May*(May 2019), 633–645.

López-Bueno, J. A., Navas-Martín, M. Á., Díaz, J., Mirón, I. J., Luna, M. Y., Sánchez-Martínez, G., Culqui, D., & Linares, C. (2021). The effect of cold waves on mortality in urban and rural areas of Madrid. *Environmental Sciences Europe*, *33*(1). https://doi.org/10.1186/s12302-021-00512-z

Petak, W. J. (1985). Emergency Management: A Challenge for Public Administration. *Public Administration Review*, *45*, 3. https://doi.org/10.2307/3134992

Peterson, T. C., Heim, R. R., Hirsch, R., Kaiser, D. P., Brooks, H., Diffenbaugh, N. S., Dole, R. M., Giovannettone, J. P., Guirguis, K., Karl, T. R., Katz, R. W., Kunkel, K., Lettenmaier, D., McCabe, G. J., Paciorek, C. J., Ryberg, K. R., Schubert, S., Silva, V. B. S., Stewart, B. C., … Wuebbles, D. (2013). Monitoring and understanding changes in heat waves, cold waves, floods, and droughts in the United States: State of knowledge. *Bulletin of the American Meteorological Society*, *94*(6), 821–834. https://doi.org/10.1175/BAMS-D-12-00066.1

Regal Ludowieg, A., Ortega, C., Bronfman, A., Rodriguez Serra, M., & Chong, M. (2022). A methodology for managing public spaces to increase access to essential goods and services by vulnerable populations during the COVID-19 pandemic. *Journal of Humanitarian Logistics and Supply Chain Management*, *12*(2), 157–181. https://doi.org/10.1108/JHLSCM-02-2021-0012

Shafapourtehrany, M., Yariyan, P., Özener, H., Pradhan, B., & Shabani, F. (2022). Evaluating the application of K-mean clustering in Earthquake vulnerability mapping of Istanbul, Turkey. *International Journal of Disaster Risk Reduction*, *79*, 103154. https://doi.org/10.1016/j.ijdrr.2022.103154

Shao, J., Wang, X., Liang, C., & Holguín-Veras, J. (2020). Research progress on deprivation costs in humanitarian logistics. *International Journal of Disaster Risk Reduction*, *42*, 101343. https://doi.org/10.1016/j.ijdrr.2019.101343

Shrivastava, P. (2003). Principles of Emergency Planning and Management. *Risk Management*, *5*(2), 67–67. https://doi.org/10.1057/palgrave.rm.8240152

Tatebe, J., & Mutch, C. (2015). Perspectives on education, children and young people in disaster risk reduction. In *International Journal of Disaster Risk Reduction* (Vol. 14, pp. 108–114). Elsevier Ltd. https://doi.org/10.1016/j.ijdrr.2015.06.011

van Wassenhove, L. N. (2006). Humanitarian aid logistics: supply chain management in high gear. *Journal of the Operational Research Society*, *57*(5), 475–489. https://doi.org/10.1057/palgrave.jors.2602125

Wright, N., Fagan, L., Lapitan, J. M., Kayano, R., Abrahams, J., Huda, Q., & Murray, V. (2020). Health Emergency and Disaster Risk Management: Five Years into Implementation of the Sendai Framework. *International Journal of Disaster Risk Science*, *11*(2), 206–217. https://doi.org/10.1007/s13753-020-00274-x

Yan, X., Xu, K., Feng, W., & Chen, J. (2021). A Rapid Prediction Model of Urban Flood Inundation in a High-Risk Area Coupling Machine Learning and Numerical Simulation Approaches. *International Journal of Disaster Risk Science*, *12*(6), 903–918. https://doi.org/10.1007/s13753-021-00384-0

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56–70. https://doi.org/10.38094/jastt1224

Zhao, J., Zhang, Q., Wang, D., Wu, W., & Yuan, R. (2022). Machine Learning-Based Evaluation of Susceptibility to Geological Hazards in the Hengduan Mountains Region, China. *International Journal of Disaster Risk Science*, *13*(2), 305–316. https://doi.org/10.1007/s13753-022-00401-w

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x