Coordenação de Pós-Graduação

PROJETO DE DISSERTAÇÃO DE MESTRADO

Dados Aluno

Nome: Renato José Quiliche Altamirano
Matrícula: 2120528
E-mail: renato.quiliche@pucp.pe
Celular: (21) 98170-2387

Dados Acadêmicos

Período em que iniciará (ou completou) o terceiro período: 14/12/2022
Data desta proposta: 14/12/2022

Professores Orientadores

De Matrícula: Adriana Leiras
De Dissertação: Adriana Leiras, Fernanda Baião

Proposta

Título da Dissertação (Provisório):

Data-driven improved pre-disaster risk reduction and preparedness interventions for recurrent countrywide clime-related disasters in Perú using Machine Learning

Data de Início: _____/_____/_____
Data de Conclusão (previsão) _____/_____/_____
Área: ( ) Gerência de Produção; ( ) Sistemas de Transporte; ( ) Finanças
Disciplinas com as quais a dissertação se relaciona (indicar com * as que tiverem sido cursadas:

- ✓ Ciencia de dados para processos de negocio*
- ✓ Estatistica e probabilidade*

Objetivo: avaliar o andamento da pesquisa do aluno, a meio-caminho entre o seu início e o seu término. Por isso ele é realizado no 3º período do curso. É baseado num documento chamado "Projeto de Dissertação", com a seguinte estrutura:

1. **Resumo da Dissertação**:

Increasing global warming harms the environment and increases the frequency of natural hazards. Greater exposure to natural hazards is increasing disaster risk. If disaster risk would not reduce, losses are expected to be more significant. Disaster risk reduction is challenging in communities affected by recurrent disasters because resources needed for response and recovery from losses are charged to communities' budgets over time. Consequently, communities have a smaller budget to invest in pre-disaster risk reduction and preparedness activities that might mitigate the magnitude of losses.

The importance of disaster risk reduction relies on building resilience against increasing natural hazards. Proactive pre-disaster activities are essential to reach this goal. Previous literature studied the determinants of disaster preparedness at the household level, performed disaster risk mathematical modeling considering economic vulnerabilities and exposure for predictive purposes, and proposed logistic strategies under uncertain conditions. However, there is scarce evidence of the validity of academic papers' conclusions in out-of-sample data. This gap traduces into a longer distance between academics and practitioners. This dissertation proposes to train a Machine Learning disaster risk classifier for households to fill this gap.

Applied Machine Learning to risk management for sudden-onset clime-related disasters and public health emergencies highlighted the importance of data science and computational intelligence for disaster risk management, as big data can be processed to extract insights in relatively short time windows. Given prior information on multiple dimensions of household vulnerability, such as economic, social, health, and geographical, the research question is: *what is the distribution of positive risk classification about being affected by an incoming disaster?* A supervised learning classifier was trained to identify at-risk households within Peruvian geographic boundaries. If a household is at risk, it represents a demand point; otherwise, it does not.

This research question was answered for three disaster types: floods, landslides, and cold waves. According to the EM-DAT disaster log, these types of disasters were selected considering their frequency and recurrence. Cold waves are localized in the Peruvian Andean region, while floods and landslides are dispersed over the country.

Train-test split strategy guarantees that the classifier trained with past data performs well with present out-of-sample data, thus overcoming the limitations of previous research. This approach creates valuable and reliable information in the context of demand uncertainty. Academics would use this information for planning humanitarian logistics, and practitioners, including critical stakeholders, would have prior knowledge of where aid should be delivered and what characteristics of households make them prone to disaster risk.

2. **Posicionamento da Dissertação no Contexto Científico e Tecnológico**: Discutir a importância do projeto de dissertação, sua motivação e a oportunidade de sua execução.

The contribution of this academic research is twofold. First, it cover a literature gap on applied machine learning to disaster risk reduction: the discussion of the model's external validity and implications for countrywide disaster risk reduction. Second, it contributes to fills the gap related to the measurement and inclusion of multidimensional vulnerability on predictive modeling for disaster risk reduction.

What motivates the project is the problem that it tries to contribute to solving. With an unavoidable increase in the frequency of natural hazards, there is an imperious need for

improved disaster risk reduction strategies. This dissertation proposes a quantitative strategy to understand risk drivers and to produce reliable predictions about disaster-specific risk distribution.
The timeliness of its implementation resides in the immediate benefit of its outcomes. Recent empirical data gathered for 2018-2020 is used to train and test a set of Machine Learning classifiers to predict households' disaster risk. The external validity of this training process is discussed within the project development. This methodology aims to guarantee that Peruvian stakeholders would benefit from optimizing resources derived from project implementation.

3. Descrição dos Objetivos do Projeto:
Quantificar e/ou qualificar as metas pretendidas.
O1: What are the characteristics of sudden onset clime-related disasters in Peru (frequency, predictability, and the magnitude of losses of floods, landslides, and cold waves)?
O2: What are the disaster risk drivers for households? What is the level of vulnerability?
O3: What would have been the impact of model implementation in the year 2020?

4. Metodologia:
Detalhar a metodologia de pesquisa adotada. Relacionar as atividades necessárias e identificar (por exemplo, com um asterisco) aquelas que possam constituir indicadores de acompanhamento da execução física do projeto. Identificar as metas já atingidas.

The methodology is quantitative. Empirical data was collected from Peruvian National Household Survey (NHS) 2018-2020. Feature engineering methods were applied to disaster risk classification. The following question was surveyed for each household (where a member responds): *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 kind of disaster. External geospatial data was used (Directory of Hydrography and Navigation, 2022; National Center for Disaster Risk Estimation, Prevention and Reduction, 2022; Dottori et al., 2015) to overcome this limitation. This data estimates the exposure area. As exposure is a necessary condition for disaster risk, all the households within the boundaries of the estimated exposure area were labeled as "households at risk" for each type of disaster. The collected and transformed data is argued to be high quality to represent the research problem empirically.
O1: Data Analytics of public data from EM-DAT disaster log. Literature review on how major Floods and Landslides affected households in Perú.
O2: Application of complete Machine Learning pipelines.

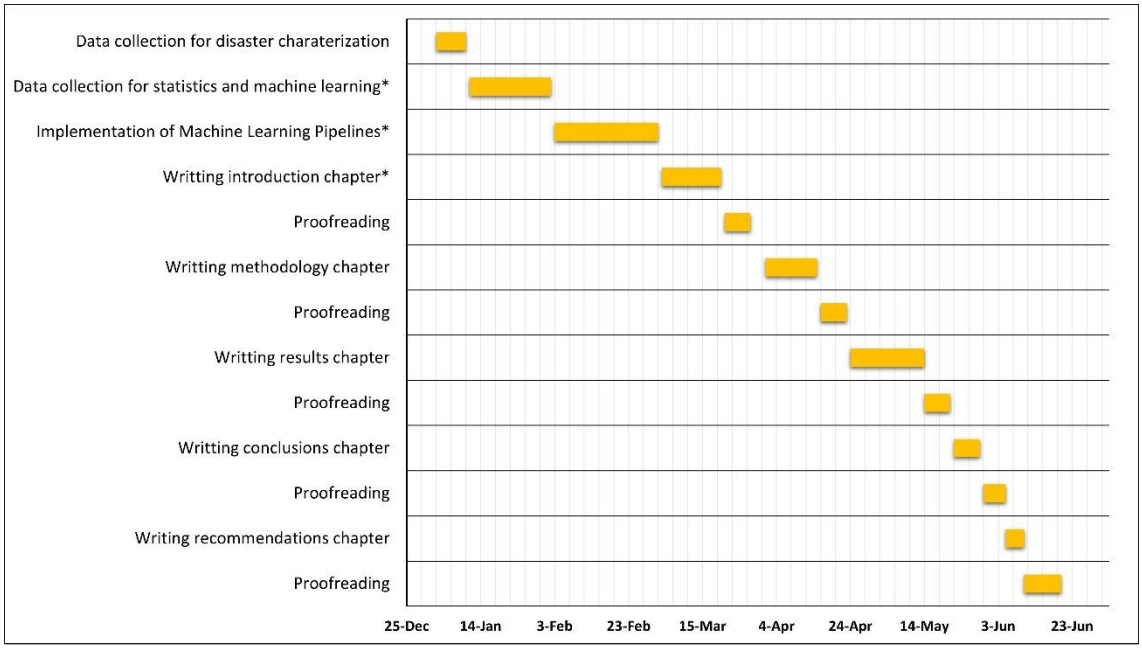
1. Feature standard scaling, feature robust scaling.
2. Supervised classifiers: Logistic Regression, XGBoost, Random Forest, Local Cascade Ensemble.
3. Performance metrics to test dataset. Geometric mean, Matthews Correlation Coefficient, ROC-AUC score, Sensitivity.
4. Results were interpreted with statistical inference for parametric models and SHAP values for ensemble non-parametric models.

O3: Experimental setting involves a train-test split strategy for the Machine Learning pipeline. Under additional assumptions regarding demand fulfillment, this allows estimating what would have been the impact of model implementation for multiple scenarios of demand fulfillment.

5. Revisão Bibliográfica:
Apresentar e analisar de forma resumida a bibliografia já revista sobre o assunto. Indicar, se for o caso, a existência de outros projetos, paralelos, que tenha relação com a presente dissertação. Relacionar a bibliografia já identificada como potencialmente relevante para o projeto e que o aluno pretenda estudar durante o andamento da dissertação.

Previous literature studied the determinants of disaster preparedness at the household level (Choi et al., 2020; Chen et al., 2022; Malmin, 2021; Lam et al., 2017), performed disaster risk mathematical modeling considering economic vulnerabilities and exposure for predictive purposes (Mors, 2010; Burton, 2010; Komolafe, 2018; Nejat and Gosh, 2016; Villaroel-Lamb, 2020), and proposed logistic strategies under uncertain conditions (Chong et al., 2019; Yang et al., 2023; Sun et al., 2022; Vaillancourt et al., 2016; Kwesi-Buor et al., 2019; Das, 2018). However, there is scarce evidence of the validity of academic papers' conclusions in out-of-sample data (Lu et al., 2021; Schumann-Bölsche, 2018).
One approach to overcome the academic-to-practice gap is to publish the results of pilot implementations of the models to assess their predictive accuracy and evaluate cost-benefit. This ideal case supposes available resources for project implementation. In the absence of such resources, an experimental evaluation of model implementation is proposed to discuss the external validity of its implications.
Another identified gap in the literature is the scarce use of multidimensional vulnerability against disasters for modeling purposes. The gross of the literature uses economic vulnerability or economic deprivation to proxy the propensity of households to be affected by a disaster (Villarroel-Lamb, 2020; Lopez-Bueno et al., 2021; Amirkhani et al., 2022). This approach is oversimplified. The advantages of adopting a multidimensional approach to disaster vulnerability were discussed in the project development: a holistic characterization of households allows Machine Learning models to draw insights from a rich set of information. Based on experimental results, this approach might overperform existing approaches.

6. Cronograma:
Desenhar um gráfico de barras e relacionar as atividades e eventos com as respectivas datas de início e conclusão previstas. Identificar as metas já atingidas.



*activities partially completed (70% of progression).

7. Outras Entidades Envolvidas no Projeto:
N.A.

8. Outros Professores e Profissionais que prestarão assistência ao aluno na execução do Projeto de Dissertação:
Paula Maçaira

9. Fontes de Financiamento Especiais para o Projeto:
This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES)

10. Congressos e Seminários:
SBPO, ENEGEP

11. Anexo Anexar as seções já escritas da dissertação (mesmo que se trate de versão preliminar)

Artigo 1 da dissertação (Puno):

A predictive assessment of households' risk against disasters caused by cold waves using machine learning

Abstract
This paper trains a household-level disaster risk classifier based on supervised machine learning algorithms for cold wave-related disasters. The households' features considered for this task proxy 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., -28°in 2003 and -35° in 2004) are recurrent and overwhelming. Two supervised learning algorithms were tested to build the classifier: 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 to maximize the model's 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. 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.

FOLHA DE ASSINATURAS

As assinaturas nesta folha formalizam a aprovação da proposta.

Aluno

Nome: Renato José Quiliche Altamirano
Matrícula: 2120528
Data: _____

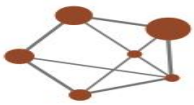
Assinatura: _____

Professor Orientador e / Coorientador

Nome: _____

Nome: _____

Assinaturas: _____



Professor Coordenador da Área

Nome: _____
Data: _____

Assinatura: _____

Professor Coordenador de Pós-Graduação

Nome: _____
Data: _____
Assinatura: _____