Climate and Disaster Risk Management for Health Systems Global Program

Climate and Disaster Risk Management for Health Systems:

A Data-Driven Artificial Intelligence Approach

Saxa 7

Emily Elwood, Paul Sweda, Mabel B. Davila, Jacob Beall, Shannon Le & Pablo Jacobo Vega

Jun 22, 2024

Executive Summary

In pursuit of the MSBA capstone project, we were sponsored by the World Bank to **identify the most critical Colombian populations lacking healthcare access and restorative services during earthquake events.** This project provides a comprehensive, actionable framework for the World Bank to evaluate Colombia's population conditions and health system impacts, specifically in the event of earthquakes as natural disasters. We establish a data-driven foundation for quantifying regions through a range of measurable risk, assisting the World Bank in synthesizing decisions for allocating aid, developing strategy, and ensuring impact in municipalities requiring the most need.

In our approach, we **employed KMeans Cluster Analysis to segment municipalities and calculated a Risk Index that assigns risk levels to population groups based on factors associated with earthquake exposure, healthcare access and socioeconomic features.** We trained machine learning models to **accurately predict this Risk Index**, capturing complex relationships between the various features. With this analysis foundation, we then developed an **Intervention Simulation algorithm** for interacting with the data. Stakeholders are able to **simulate various intervention scenarios—such as increasing healthcare access or improving access to sanitation and to resultantly observe how these changes impact the Risk Index.** The simulation provides valuable insights into which interventions are likely to reduce vulnerability most effectively, guiding targeted actions and optimizing resource distribution for maximum impact. Populations at-risk in the context of this project is defined by the degree to which communities are at an increased risk of harm or adverse health outcomes due to a combination of geographic factors, health system inadequacies, infrastructure conditions, resource recovery shortages, pre-existing health conditions, and other characteristics that compound sensitivity to hazards upon exposure. Decision-makers will be able to test possible funding or improvement strategies, and view predictable outcomes immediately.

All analytical findings followed by the simulation algorithm will assist in bridging data insights with viable suggestions to ultimately reduce critical concerns, and to provide the means for building climate-resilient health systems, mitigating disaster risks, and improving public health outcomes. Additionally, beyond improvement, this targeted approach will help strengthen the national systems to be equipped for disaster challenges through clear, evidence-based support for policy proposals.

Methodology

The methodology consists of data preparation, cluster analysis, model training, and intervention simulation. Each listed component will be further elaborated on.

1. **Data Collection and Preparation:** Gathered data on healthcare accessibility, socioeconomic factors, geospatial and population demographics. Data was cleansed, preprocessed, and normalized where necessary to ensure consistency and readiness for analysis.
2. **Cluster Analysis and Risk Index:** The proprietary preprocessing (GFDRR preprocessing) framework established by the Climate and Disaster Risk Management for Health Systems Global Program calculated the exposure, risk index and segmented municipalities using KMeans based on vulnerability characteristics established during this analysis. Clustering allowed us to group municipalities with similar risk profiles, highlighting which municipalities should be prioritized for evaluation on potential aid or process improvement based on their current risk patterns.
3. **Risk Index Prediction using Machine Learning:** Selected the best performing ML model to predict a new Risk Index evaluating each municipality's effect after applying intervention scenarios. This index was predicted using the *Gradient Boosting Regressor algorithm.* The model was trained on historical vulnerability indicators, and hyperparameters were tuned to optimize performance, with *RMSE* as the primary evaluation metric.
4. **Intervention Simulation:** To support decision-making, we developed an *intervention simulation algorithm.* This tool allows stakeholders to test hypothetical scenarios by adjusting key indicators (e.g., reducing the need for protection resources in a municipality) and observing the *impact on the Risk Index*. By applying simulated interventions across all municipalities, we can compare predicted outcomes to the baseline Risk Index, identifying the most effective interventions for reducing risk.
5. **Export and Visualization of Results:** Simulation predictions, including actual and predicted Risk Index values for three intervention scenarios, were exported to GEOJSON and CSV files for future analysis. These formats were chosen to facilitate future analysis and integration into a variety of analytical tools and workflows. mock-up visualizations were developed to represent the data dynamically. These visualizations were designed to display clusters of municipalities and the outcomes of the simulation on an interactive map. By overlaying the risk scores and intervention impacts on geographical boundaries, stakeholders can easily identify patterns, high-risk areas, and the effectiveness of proposed interventions.

Data Collection and Preparation

Beginning with a crucial, preliminary data collection, we conducted a multidimensional process to research, assess and combine data sources composed of Colombia's country profile in order to integrate an adequate dataset to study. In combining various open source datasets, we were able to create a comprehensive, cumulative socio-demographic profile of Colombia to begin structuring a risk model. Datasets included spatial distribution of 2020 population, GIS OpenStreetMap (points of interest, roads, etc.), health facilities locations, global earthquake hazards (probabilistic seismic hazard assessment at a global level). Additionally, we included 'risk features' datasets for the 2018 Housing Census, 2024 Housing Subsidies Allocated, 2024 People In Need And Severity and the 2023 4W which provides key information regarding which organizations (Who) are carrying out which activities (What) in which locations (Where) in which period (When). Refer to the data catalog for detailed information on each dataset.

Preparing the data for analysis involved standard data cleaning, translating from Spanish to English language, renaming columns, assigning correct data types, removing columns with more than 90% of missing values. Feature engineering was required for some of the datasets:

- *Data_Cleaning_345_W_dataset.ipynb*: many organizations had multiple response activities listed per Municipality. We summed all the numeric columns excluding the boolean columns. The boolean columns and non-numeric columns retained the most common value. This approach ensures that the most common 4W_CLUSTERSECTOR is retained when aggregating records for each ADM2_C. If there is a tie in the most frequent value in the group, then the first mode is selected.
- *Data_Cleaning_Subsidios_De_Vivienda_Asignados.ipynb:* a dataset of 80,567 records and 9 columns from 2003 to 2024 listing several housing programs that were approved and not approved for different reasons. In this case, we filtered records for the year 2024 (down to 4,114 records) then applied a logic to create a new boolean column HOUSING_IS_RESILIENT where benchmarks were calculated using the median to derive a threshold for each column ASSIGNED_VALUE and HOMES. The threshold is used to determine if a municipality is over- or under-served compared to the median values. Then grouped all the records of the same municipality to reflect that 276 (rows) municipalities had a record of receiving housing subsidiaries. As the final feature engineering step for this individual dataset, we joined the 276 municipalities with the missing municipalities to complete the total of 1,122 municipalities where we replaced nulls with 0s to reflect these municipalities did not receive housing subsidiaries, therefore were considered NOT resilient. Resiliency in our logic is defined as the status of the municipality (yes = 1 or no = 0) if any low-income households, classified by Colombia's SISBEN, System of Identification of Social Program Beneficiaries, within the municipality received financial assistance to purchase or build a home meeting the minimum quality and safety standards required by the government (Vélez et al., 1999). With this logic, the dataset of 9 columns and approximately 80k rows was transformed to be useful in the analysis as one column with 1,122 rows and merged to the main dataset *risk_index_features.csv* using Colombia's subnational boundaries code, ADM2_C, as the unique identifier. This .csv dataset (1,122 rows and 25 columns) feeds the proprietary preprocessing (GFDRR preprocessing) framework established by the Climate and Disaster Risk Management for Health Systems Global Program.

The GFDRR preprocessing framework included extracting and transforming the cleaned individual datasets, followed by completing the *thresholds_file.xlsx* by entering the corresponding calculated Ideal Value for each risk feature. In doing so, we performed standardization, calculating exposure per administrative region and processing hazards for every earthquake return period to establish the Risk Index. Return periods 475, 975, 1500, and 2475 are standards for assessing seismic risk. Additionally, the preprocessing employed a KMeans Cluster Analysis to segment municipalities, highlighting those with the highest exposure levels that should be prioritized for aid. **We decided to not treat outliers. Our primary goal is risk identification and mitigation, we believe that these outliers reflect true Real-World risks, keeping all outliers is beneficial.** For example, in the Boxplot and Histogram for "ELECTRICITY" [Figure A0], there are several data points below the lower whisker, representing municipalities with significantly more access to electricity. These outliers align with the left tail of the histogram and are likely the ones causing the left skew. The approach of not treating outliers ensures that the model remains responsive to extreme cases, which is the primary focus in risk assessment. Removing or modifying true outlier cases could introduce bias by artificially lowering or increasing the risk index, leading to underestimations and missed opportunities for necessary interventions. The preprocessing framework resulted in *output risk datasets* for each return period. **Our main focus was the output risk data for the return period 475, indicating that a seismic event of a certain magnitude has a 10% probability of occurring within a 50-year timeframe, meaning that on average, is expected to happen once every 475 years.**

We incorporated 4W_SECTORCLUSTER, a categorical column, to the *output risk data* to enrich simulation scenarios. This approach allowed us to refine the analysis by addressing the current assistance programs in the country, filling the gaps in the initial clustering during the preprocessing stage. This approach also makes it possible to add newly available data, making the model more adaptable to real-world changes. In a future phase for the project, this column could be considered in the municipality re-clustering and risk index calculation during the preprocessing stage. Since this column is categorical, it was critical to convert it to a numerical data type by **One-Hot encoding** to:

> 4W_SECTORCLUSTER_Food_Security_and_Nutrition, 4W_SECTORCLUSTER_Health,
> 4W_SECTORCLUSTER_NO_ASSISTING_PROJECT, 4W_SECTORCLUSTER_Protection,
> 4W_SECTORCLUSTER_Socioeconomic_Early_Recovery,
> 4W_SECTORCLUSTER_Water_Sanitation_and_Hygiene and,
> 4W_SECTORCLUSTER_Emergency_Education.

Followed by **splitting the data into Train and Test with a 80-20 split.** After running the base model, Multiple Linear Regression, we applied the forward_selection function. **The forward selection approach takes the feature matrix *X* and target *y* as inputs and iteratively adds the feature that minimizes the Akaike Information Criterion (AIC)**, an estimator of prediction error**.** The feature that most reduces the AIC is added to the model. This process repeats until no further improvement is observed.

Every dataset cleaning process and the *Output Risk Analysis* was done individually using a separate jupyter notebook file. These notebooks are located in the project folder:

> *Risk_index_Col_earthquake-main\notebook*

Approach & Assumptions

Open-source datasets offered data in different timeframes where we opted to select the most recent data available. Some **assumptions were made on each of the risk features datasets.** For example, in the *4W* dataset the column 4W_BENEFICIARY_TOTAL_PEOPLE is the sum of the columns 4W_BENEFICIARY_GIRL, 4W_BENEFICIARY_BOY, 4W_BENEFICIARY_FEMALE, 4W_BENEFICIARY_MALE. In some cases, the dataset specified the distribution of the beneficiaries and in some records, it had 0 across all 4 columns only with the Total People column with a true numerical value. We applied the following rule to replace 0s with non-zero values that are helpful to count the beneficiaries in each category. If the four columns had 0s and the

4W_BENEFICIARY_TOTAL_PEOPLE had a value greater than 1, then replace the 0s with the result from dividing the total value by four. For example, if total people is 11, the beneficiaries girl, boy, female and male will have a value of 2.25. This approach helped us in finding the Ideal Value using the mean for each of those columns.

It is important to note that the GFDRR preprocessing framework excluded the 4W_SECTORCLUSTER since it only takes numerical type data. Instead, after the GFDRR preprocessing, we used the output for *risk_index\earthquakes_period_1-475* and left-joined it to the 4W dataset to bring in the 4W_SECTORCLUSTER and then one-hot encode it.

Exploratory Data Analysis

During the Multivariate Analysis, we used the Spearman correlation method when computing the correlation matrix. **Spearman method is rank-based and is suitable for capturing non-linear relationships.** Features with a correlation above 0.7 and below -0.7 were removed. In maximizing positive impact from the Correlation Matrix, we accounted for insights that emphasize the correlations between service variables in order to design integrated improvement plans that upgrade multiple services *simultaneously*. We suggest leveraging the positive correlations among social services to bundle child protection, gender violence prevention, and general protection in strategic locations. We also emphasize increasing beneficiary coverage in high-risk areas to potentially reduce the risk index, based on the observed negative correlation.

Visualizing 4W_SECTORCLUSTER we can see most of the humanitarian response activities were dedicated for Protection, then Food, Security and Nutrition and then Health [Figure A1].

Understanding the target variable provides a foundation for prioritizing areas for aid and refining predictive models to maximize their practical application. Risk Index, the target variable, represents the at-risk population of Colombian municipalities. Based on the histogram ‘Risk Index Distribution’ [Figure A2], Risk Index follows a slightly skewed distribution with a peak range between 0.15 to 0.20. This suggests the majority of municipalities fall within the moderate risk range indicating most municipalities share similar levels of exposure. The long tail on the right suggests the presence of municipalities with significantly higher risk, the outliers with unique conditions contributing to heightened exposure.

Model Fitting

Starting with a base model, we performed a **Multiple Linear Regression (MLR)** with diagnostics plots. **This model provided a strong baseline for predicting the Risk Index.** The significant predictors highlight important relationships that can guide strategic decisions for intervention. However, this model failed to meet the assumptions of Linearity, Homoscedasticity and Normality. Without meeting these criteria, the model's predictions could be unreliable and misleading. **Thus, this performed MLR confirmed that instead we need to consider non-linear models to capture the non-linear relationships.**

The **Support Vector Machine (SVM)** is a powerful machine learning algorithm that can be used for regression and classification tasks. **This non-parametric model works well with non-linear and high-dimensional data.** We utilized a GridSearch with 10 folds for cross-validation to find the following best parameters: `{'C': 1, 'epsilon': 0.01, 'gamma': 0.01, 'kernel': 'rbf'}`. The R^2 of 0.9227 means the variance in the target variable is explained by the model by 92.3%. Although the R^2 is high, *there is a risk of overfitting*. The MSE (0.0001) is very low, which implies that the model's predictions are close to the actual values. The RMSE of 0.012

suggests that, on average, the model's predictions are off by only 0.012 units. This is relatively small and points to a model that is performing with high accuracy. The MAE indicates that the average absolute difference between the actual and predicted values is 0.009. This means the model's predictions are quite precise, with only a small average error per prediction.

The **Neural Network (NN)** model is a complex modeling technique consisting of function approximators, mapping inputs to outputs. The computational units within the model making are called neurons. Approximately 84.98% of the variance in the Risk Index is explained by the model. This is a strong indicator of good predictive power. The RMSE at 0.0165 suggests that the average prediction error is about 0.0165 units from the actual values, which implies that the model's predictions are quite close to the true values. The R^2 shows that the model is performing well with strong predictive capabilities. However, the **slight deviations in the diagnostic plots, such as the Q-Q plot, suggest that there may be some non-linearity or outliers that are not fully captured by the model.**

The **Gradient Boosting Regressor (GBM)** is well-suited for capturing non-linear relationships between features due to its use of decision trees as base models. GBM does not require feature scaling and is suitable for complex data. **This model does not require scaled data because it uses distance-based calculations in which splits nodes based on feature values. Outliers that are genuine indicators of extreme risk were kept.** The best parameters for GBM are: `{'learning_rate': 0.05, 'max_depth': 3, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 300}`. **The high R^2 value of 92.6% indicates that the GBM model explains a large portion of the variance in the Risk Index, making it a strong model for prediction.** The model's diagnostic plot 'Residuals vs. Predicted Values' suggests there are no clear clusters or patterns. It means that the model captures the relationship between features and the target variable well [Figure B1]. The 'Q-Q Plot for Residuals' reveals there are slight deviations at the tails, indicating some minor issues with normality. This is generally acceptable in GBM models, as they do not rely heavily on normally distributed residuals [Figure B2]. In the 'Histogram of Residuals', the residuals appear to be centered around zero, with a distribution close to normal [Figure B3]. In the 'Scale-Location Plot', the residuals appear to be fairly homoscedastic (constant variance) across the range of fitted values, with a slight concentration around the center. **The diagnostics indicate that the tuned GBM model performs well and is a strong predictor of Risk Index.** There are no major issues with bias, normality, or homoscedasticity in the residuals. **The model metrics confirm that this GBM model is well-suited for this prediction task.**

Model Comparison

The RMSE is more robust with maintaining lower values despite the presence of outliers, therefore we are using RMSE as our main error metric to compare our models. Both GBM and SVM are well-suited for capturing non-linear relationships. **The Tuned GBM achieved the best performance with the lowest RMSE (0.0116) and the highest R^2 (0.926).** This suggests it has the best predictive power among the models and thus, we prioritized using it for insights and interventions. SVM came in close to the Tuned GBM with an RMSE of 0.0118 and an R^2 of 0.923, indicating it also has good predictive power in handling the variability in the data well. This could be used as a secondary analysis option for future extensive research. Baseline Multiple Linear Regression performed reasonably well, with an RMSE of 0.0134 and an R^2 of 0.900. However, it does not capture the data's non-linearity as effectively as the more complex models. Neural Network had the highest RMSE (0.0165) and the lowest R^2 (0.850), indicating it struggled more with this dataset, possibly due to its sensitivity to the outliers.

Gradient Boosting Regressor Model Insight

Permutation Importance is a technique used to assess the importance of individual features within a model by randomly shuffling the values of a single feature and observing how much the model's performance degrades, essentially measuring how much the model relies on that feature to make predictions; **a larger drop in performance indicates a more important feature**. In analyzing these findings, we are able to understand how much the model depends on the feature. (scikit-learn, *4.2. permutation feature importance*)

The range of values from negative to positive in the 'Permutation Feature Importance' plot indicates that some features, when shuffled, have a strong negative impact on the model's performance, while others show a positive or minimal impact [Figure C1]. In complex models like gradient boosting, features can interact in ways that cause these mixed effects. For example, some features may only be useful in combination with others, so shuffling them in isolation has a different effect on model performance than shuffling more independently valuable features. This results in some features showing a clear importance (positive or negative), while others have a smaller or inconsistent impact on the prediction accuracy. This variety in importance values reflects the complex and sometimes non-linear relationships among the features in predicting the target variable. **Features PERC_PROTECTION and PERC_HEALTH have the most negative impact when shuffled. This might indicate that while they influence the Risk Index, their contribution might sometimes be indirect or interact with other features in ways that do not always improve predictive accuracy. In this case, there is a chance that their real predictive power is tied to their interaction with each other.** Same thought process applies for PERC_MINE_ACTION or ELECTRICITY. The 4W_BENEFICIARY_TOTAL_PEOPLE feature is the most positive influential predictor in the model. This suggests that municipalities with higher or lower counts of direct or non-direct beneficiaries are influencing the risk assessment.

The 'Partial Dependence Plots' offer a comprehensive understanding of how individual features impact risk predictions, guiding targeted policy and resource allocation. **A partial dependence plot (PDP) is a visualization tool that shows how a feature or features affect a machine learning model's predictions by helping understand the relationship between a feature and the model's predictions, while holding all other features constant.** Whether the relationship between the feature and the target is linear, monotonic, or more complex. The feature importance tells how important the feature is to the model's predictions while the direction of influence is the direction in which the feature influences the prediction [Figure D1].

- **4W_BENEFICIARY_TOTAL_PEOPLE**: The flat line suggests that changes in the number of total beneficiaries in the program do not significantly influence the predicted risk possibly because other factors like the type of assistance or health services might carry more weight.
- **exposure_health_care_hospital**: The linear trend suggests that increased exposure or access to healthcare services consistently influences the Risk Index positively, with no sign of diminishing returns within the range shown. This implies that healthcare exposure is a critical and steady contributor to risk levels, and increasing healthcare access may directly relate to better risk outcomes.
- **HOUSING_IS_RESILIENT**: As HOUSING_IS_RESILIENT increases, the Risk Index decreases substantially. Municipalities with stronger, disaster-resistant housing tend to have lower exposure.
- **4W_BENEFICIARY_AFROCOLOMBIANO**: The Risk Index tends to decrease slightly as the value of 4W_BENEFICIARY_AFROCOLOMBIAN increases indicating that increasing aid or support to Afro-Colombian populations might have a small, but positive effect in reducing vulnerability. The effect is subtle, suggesting this feature may not be a dominant driver of the Risk Index.

- **4W_BENEFICIARY_INDIGENOUS**: Providing aid to Indigenous populations appears to have a modest effect in reducing risk.
- **NATURALGAS**: The relationship between access to natural gas and the Risk Index is non-linear and somewhat complex. There is significant variation with some instances leading to an increase or decrease in the Risk Index. This feature's non-linear behavior may indicate interactions with other variables. While access to natural gas might improve living conditions, the data suggests that it alone does not consistently reduce risk and might even contribute to higher risk in some cases.

Intervention Simulation

The Intervention Simulation uses our trained Gradient Boosting Model as the predictive approach to assess how changes (interventions) in key vulnerability factors could reduce the risk index for municipalities. This approach enables stakeholders to simulate various intervention scenarios—such as decreasing the need for healthcare resources or improving infrastructure—and observe how these changes impact the Risk Index. The simulation provides valuable insights into which interventions are likely to reduce risk most effectively, guiding targeted actions and optimizing resource distribution for maximum impact. As an example, the following are 3 intervention scenarios

1. Lower the need of healthcare resources by 40%, increase access to electricity by 20%
2. Lower the need of healthcare resources by 40%, increase the access to electricity by 20%, and lower the need of protection resources by 10%
3. Lower the need of healthcare resources by 40%, lower need of protection resources by 10%

Results and Recommendations

The Gradient Boosting Model (GBM) performed well in predicting the Risk Index, achieving a low RMSE of 0.0116 and high R^2 score of approximately 0.93. This suggests that the model captures a significant portion of the variance in the Risk Index based on the features provided. The Permutation Feature Importance plot highlights that PERC_PROTECTION, PERC_HEALTH, and PERC_MINE_ACTION have the most substantial impact on the model's predictions, followed by features such as ELECTRICITY and PERC_EMERGENCY_EDUCATION. **The presence of negative importance values for certain features indicates complex interactions in the model, where creating interactions, these features could have varying effects.** For example, PERC_PROTECTION and PERC_HEALTH may affect outcomes differently when combined with other features. The Partial Dependence Plots indicated 4W_BENEFICIARY_TOTAL_PEOPLE show a minimal variation in Risk Index, possibly its influence is limited to high population areas. NATURALGAS shows a nonlinear relationship, possibly due to interactions with other features.

Given the high importance of PERC_PROTECTION and PERC_HEALTH, *increasing support in these areas is likely to have the most significant positive impact on reducing the Risk Index*. Initiatives in these sectors should be prioritized, especially in areas with low current values. PERC_MINE_ACTION also has a notable impact on risk reduction. Interventions in mine clearance and education on landmine risks can be targeted to improve safety and reduce exposure in high-risk areas. Enhancing access to ELECTRICITY and

PERC_EMERGENCY_EDUCATION could contribute to improved resilience, therefore infrastructure development in electricity and emergency preparedness programs should be emphasized. Interventions should consider increasing support or tracking in areas with a high number of direct and indirect beneficiaries or large populations such as 4W_BENEFICIARY_TOTAL_PEOPLE has an effect on the Risk Index. The insights from the Correlation Plot suggest the benefit of integrated improvement plans that upgrade multiple services simultaneously. Leveraging the positive correlations among social services, the suggestion is to bundle child protection, gender violence prevention, and general protection in strategic locations. Focusing resources on the features with the highest positive impact according to the Permutation Feature Importance, Partial Dependence Plots and Correlation Plot, interventions can be targeted effectively to maximize reductions in the Risk Index. Additionally, we suggest a future analysis to investigate further the non-linear relationships and interactions, like NATURALGAS and HOUSING_IS_RESILIENT, to uncover hidden insights that may guide more effective resource allocation.

Conclusion

This project successfully identified and prioritized municipalities in Colombia that are most at-risk in accessing healthcare services during earthquake events. By developing a comprehensive Risk Index, municipalities were categorized based on their level of risk using KMeans Cluster Analysis, enabling targeted intervention planning.

Machine learning models, such as the Gradient Boosting Model (GBM), were trained to predict the Risk Index with high accuracy, indicating the model's strong ability to capture complex relationships. The Intervention Simulation algorithm further enhanced decision-making by allowing stakeholders to test "what-if" scenarios and evaluate the impact of targeted interventions on the Risk Index.

Key features such as PERC_PROTECTION, PERC_HEALTH, and PERC_MINE_ACTION were identified as having the most significant influence on risk exposure. These insights guided recommendations for prioritizing investments in healthcare access, protection programs, and infrastructure improvements like electricity and sanitation.

By combining advanced data analysis with predictive modeling and scenario testing, this project provides the World Bank with a robust, data-driven framework to optimize resource allocation and build resilience in Colombia's healthcare system. The outcomes ensure that aid is directed to areas where it can have the greatest impact, supporting Colombia's broader climate and disaster risk management goals.

Post Presentation Questions and Clarifications

Distinguishing between the impact of variables with the same rate of change: Mathematical distance is retained when normalizing each variable independently. By applying methods like min-max scaling or z-score standardization to each variable separately, the relative distances between data points within that variable's distribution remain proportional. This ensures that the relationships and variability in the data are preserved. Normalizing independently also prevents distortions caused by differing scales or units across variables, allowing for fair comparisons without one variable dominating the analysis due to its raw scale. This approach maintains the integrity of the data while enabling meaningful and balanced contributions to the overall model.

Confidence interval and sensitivity analysis: On a relative index our example municipality was at a 99 percentile in terms of risk score after intervention 1 it was 78th, intervention 2 was in the 63rd, and intervention 3 it was in the 79th percentile. We believe this more accurately contextualizes the change in risk score.

Appendix

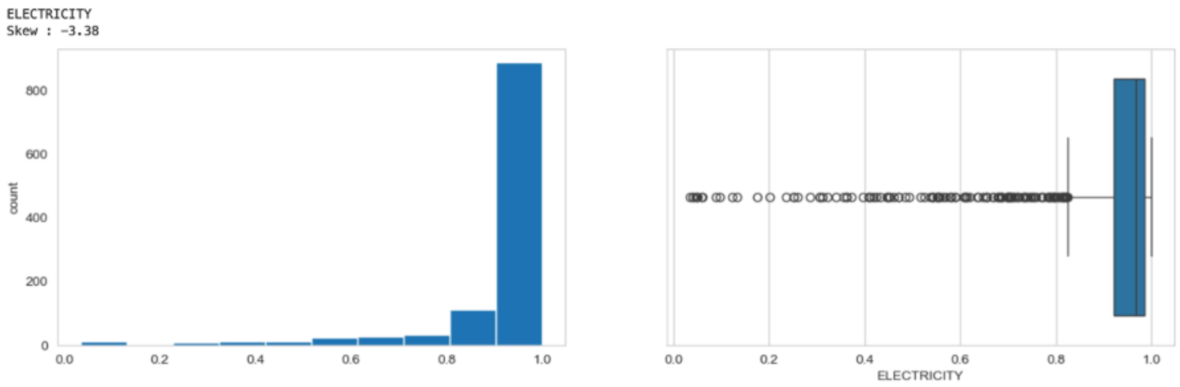

Figure A0.
Boxplot and Histogram for "ELECTRICITY".

A large number of observations fall near 1.0, implying that most municipalities have the most need to access electricity.

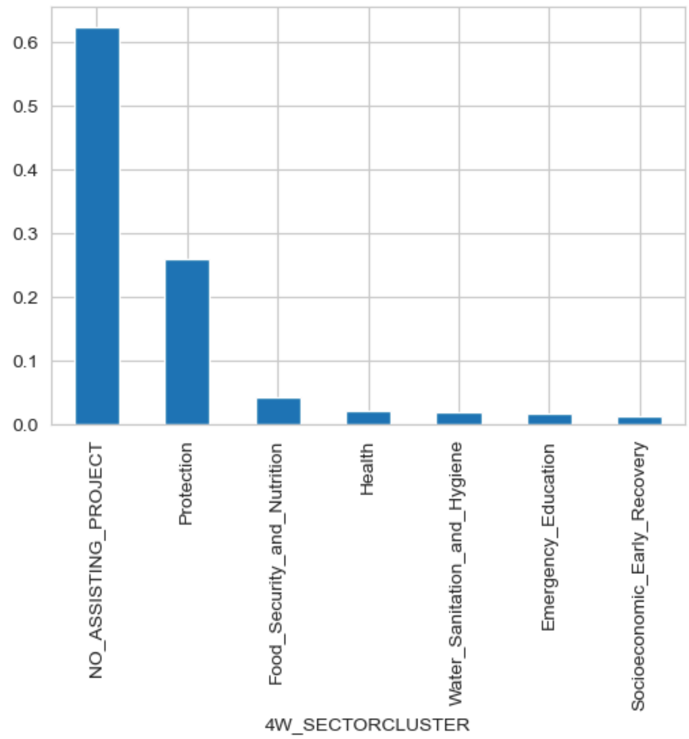

Figure A1.
Humanitarian Response.

4W_SECTORCLUSTER suggests most of the humanitarian response activities were dedicated to Protection, then Food, Security and Nutrition and then Health.

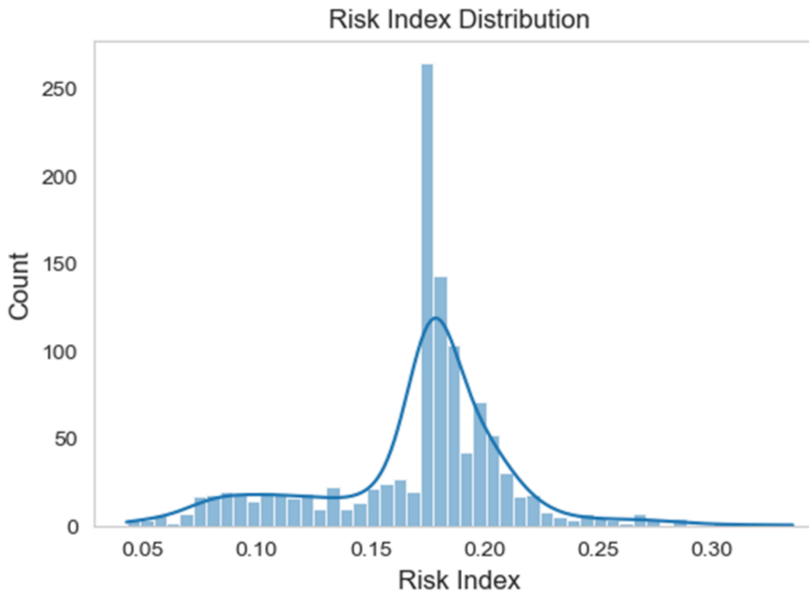

Figure A2.
Risk Index Distribution

Most areas share similar levels of risk.

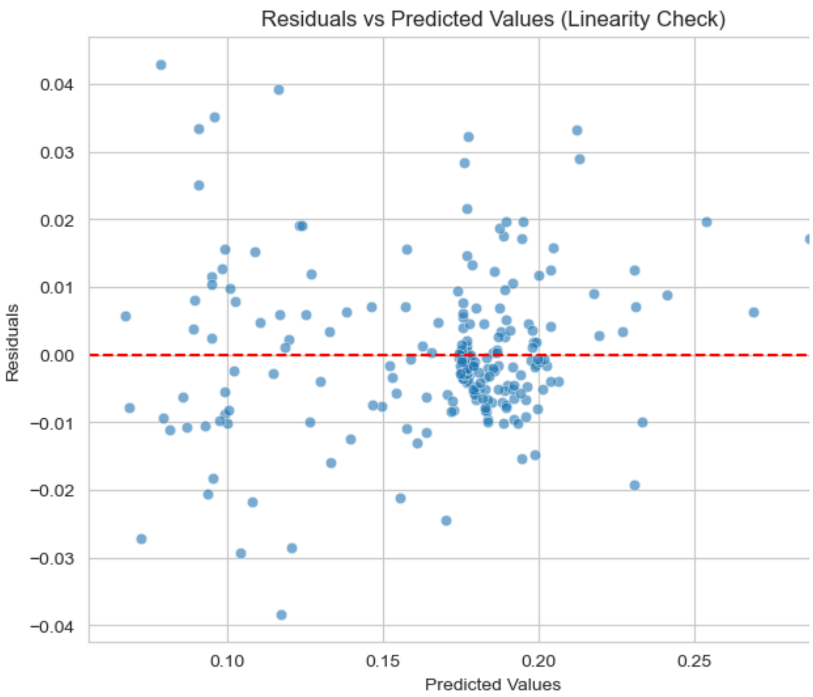

Figure B1.
Residuals vs. Predicted Values for the GBM.

The residuals (differences between actual and predicted values) are scattered around the horizontal axis at 0, without any clear pattern, indicating that the model predictions are, on average, unbiased and that the model captures the data relationship well.

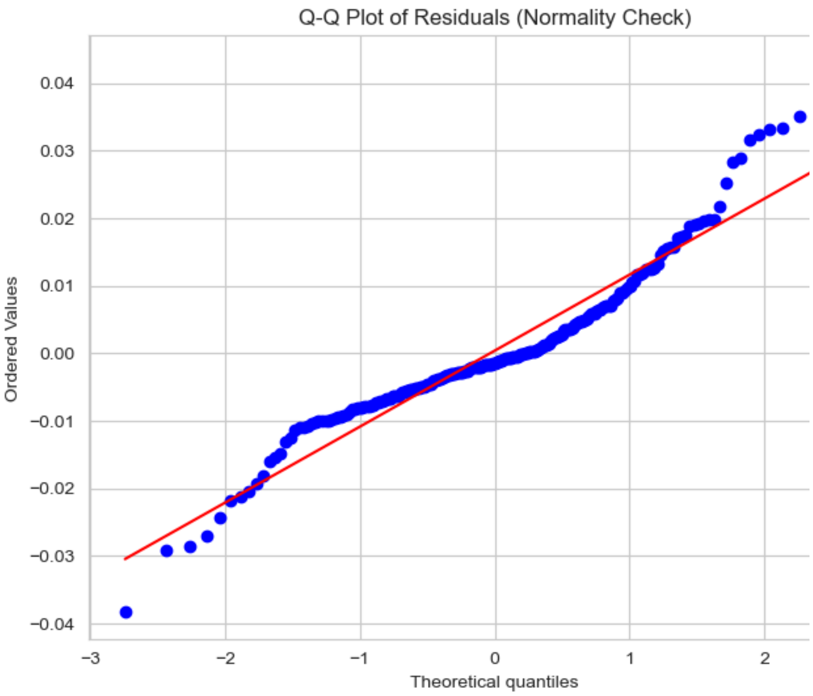

Figure B2.
Q-Q Plot of Residuals for the model GBM.

The residuals closely align with the red diagonal line in the central portion of the plot, indicating that the residuals are approximately normally distributed around the mean for most data points.

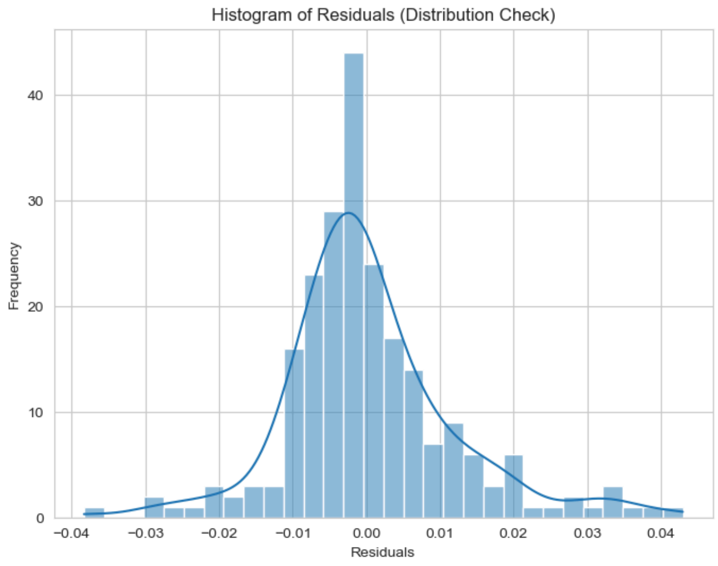

Figure B3.
Histogram of Residuals (Distribution Check).

The residuals appear to be centered around zero, with a distribution close to normal.

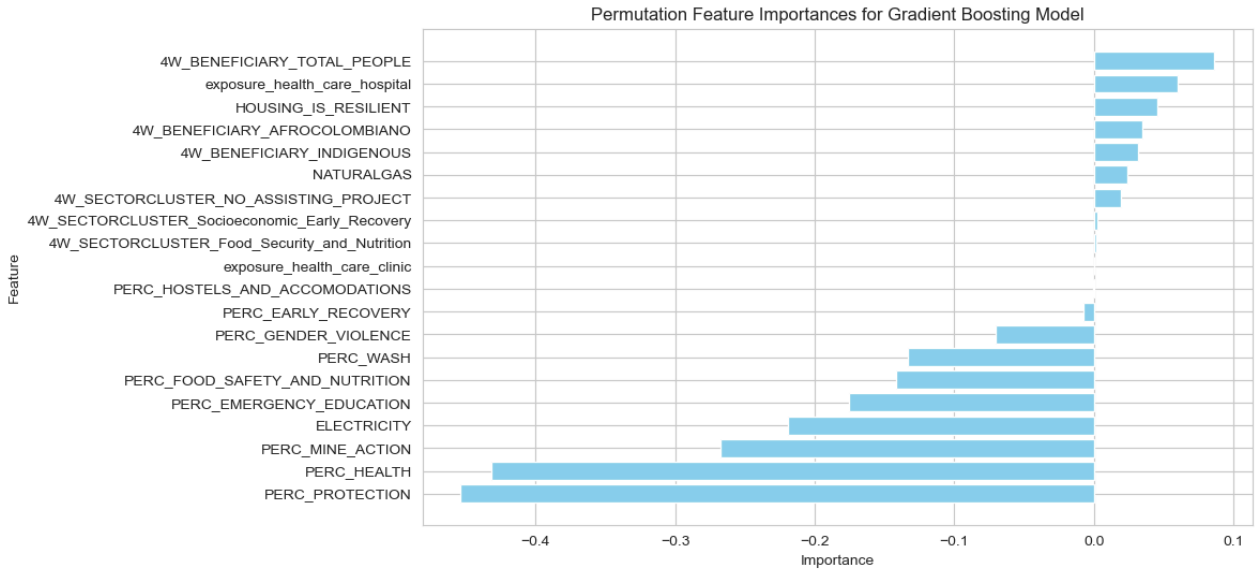

Figure C1. Permutation Feature Importance.
PERC_PROTECTION and PERC_HEALTH have the most negative impact. This might indicate that while they influence the Risk Index, their contribution might sometimes be indirect or interact with other features in ways that do not always improve predictive accuracy. In this case, there is a chance that their real predictive power is tied to their interaction with each other. Same thought process applies for PERC_MINE_ACTION and ELECTRICITY.

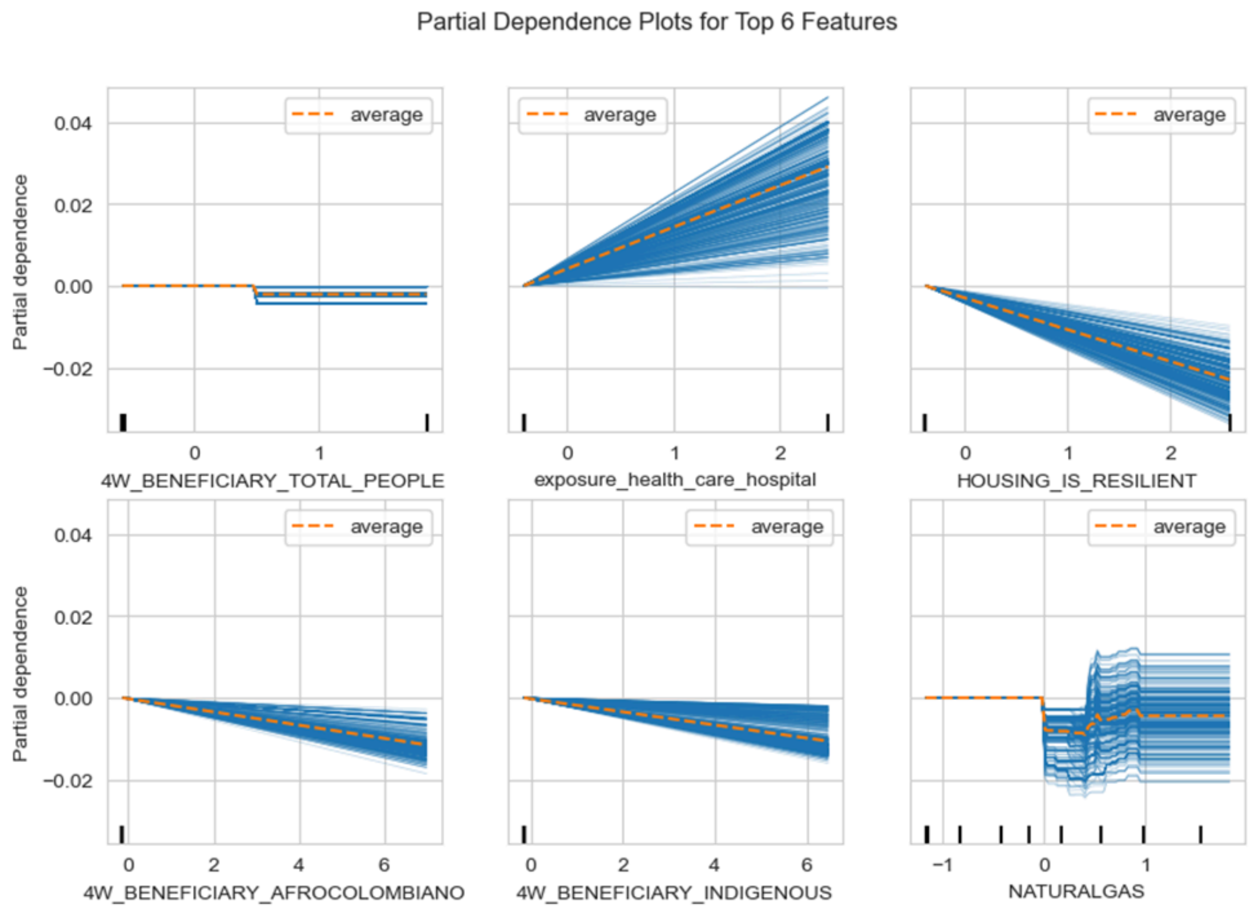

Figure D1. Partial Dependence Plots.
The feature importance tells how important the feature is to the model's predictions while the direction of influence is the direction in which the feature influences the prediction.

References

World Bank. GFDRR Annual Report 23 - Bringing Resilience to Scale (English). Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/099836402122412500/IDU1296966181302414785188c41e3492095ce66

Think hazard - Colombia - earthquake. (n.d.). https://thinkhazard.org/en/report/57-colombia/EQ

Geometric objects - spatial data model¶. Geometric objects - Spatial data model - Intro to Python GIS CSC documentation. (n.d.). https://automating-gis-processes.github.io/CSC/notebooks/L1/geometric-objects.html

Centers for Disease Control and Prevention. (2024, June 14). CDC/ATSDR social vulnerability index (CDC/ATSDR SVI). CDC/ATSDR Social Vulnerability Index (CDC/ATSDR SVI): Overview. https://www.atsdr.cdc.gov/placeandhealth/svi/index.html

Drakes, O. (n.d.). Multi-scale Multi-hazard Social Vulnerability. https://www.youtube.com/watch?v=K6wwDYAzPgI

Vélez, C. E., Castaño, E., & Deutsch, R. (1999, May). An Economic Interpretation of Targeting Systems for Social Programs: The Case of Colombia's SISBEN.

Saavedra, P. F. (2024, August 13). Afectados por el Fenómeno de La Niña 2021-2022 Podrán reclamar hasta $500.000: Esto es lo que deben hacer. infobae. https://www.infobae.com/colombia/2024/08/13/nuevo-ciclo-de-pagos-del-programa-de-ayuda-economica-para-afectados-por-la-nina-2021-2022/

RBF SVM Parameters. scikit. (n.d.-c). https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

Mlpregressor. scikit. (n.d.). https://scikit-learn.org/dev/modules/generated/sklearn.neural_network.MLPRegressor.html

Gradientboostingregressor. scikit. (n.d.-a). https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

4.2. permutation feature importance¶. scikit. (n.d.). https://scikit-learn.org/0.24/modules/permutation_importance.html#:~:text=The%20permutation%20feature%20importance%20is,model%20depends%20on%20the%20feature

Mi Casa Ya. Minvivienda. (n.d.). https://www.minvivienda.gov.co/viceministerio-de-vivienda/mi-casa-ya

OpenAI. (2024). ChatGPT conversation on data-driven analysis for prioritizing vulnerable Colombian populations in healthcare during earthquake events. ChatGPT model, version GPT-4-turbo. Retrieved from https://chat.openai.com/

Data Catalog: https://drive.google.com/file/d/1GEs-ck6T1qxRYX1Wo-OI0UTI2O-1kC4n/view