EARLY RISK ASSESSMENT OF COMMUNITIES PRONE TO FLOODING IN COLOMBIA

Team 30: DATA TIGERS | David Bonilla, Edwin Cifuentes, Daniel Córdoba, Javier Fernández, Evelyn Guerrero, Natalia Hoyos, Juan Villanueva.

HIGHLIGHTS

- It's a complete, fast, and easy-to-use application trained with data up until 2020 and tested on data from 2021. After iterating with different approaches and optimizing our pipeline, we got the best model to feed the app.
- Now you can have an early warning and reduce the number of casualties and damages suffered by many communities.

BACKGROUND

Rainy seasons usually involve floodings that cause infrastructure loss, agricultural assets damage, and food security issues. In Colombia there are almost 600 floods annually and they usually leave more than 2 million people affected. Although UNGRD has implemented early warning systems to emit lifesaving alerts in these cases, the current system does not allow performing and spreading predictions of floodings in high-risk areas.

Accumulated rainfall is the main cause of flood emergencies

Floods are the most recurrent emergencies among 20 types

Floods are the leading cause of death from hydrometeorological phenomena

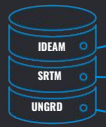
OBJECTIVE

Creating a model that predict floodings in the high-risk areas in Colombia, based on available GIS open data to enhance public entities resource allocation processes and perform preventive actions more effectively, improving the current Natural Disasters Response system capabilities.

DATA

The analysis period was limited to 5 years, for the municipalities with the highest historical records of floodings, using data from:

Sources



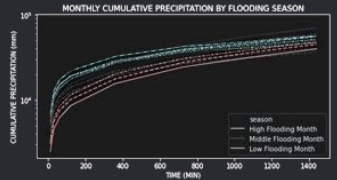
Basic data transformation

- To aggregate each variable to get the daily average, minimum, maximum and standard deviation.
- To load the datasets to BigQuery and performed the aggregation through a query.
- To merge all datasets through divipola codes to filter and clean non available data.

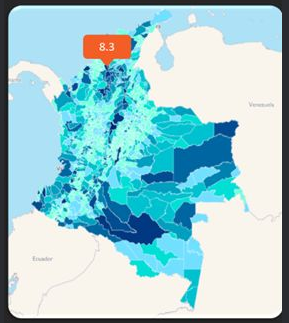
Limitation: The south western of Colombia is significantly underrepresented.

PROCESSING

We had more than 100 times less samples in the flood class than the control making our data heavily imbalanced. So we had to oversampled the minority classes until the classes were balanced.



We used 5-fold cross-validation and trained multiple models with the data up until 2020 and tested with data from 2021. With the strength of scikit learn pipelines and grid search.



ARCHITECTURE

The DASH platform was selected for the FrontEnd supported by PostgreSQL and Google Bigquery. This platform has an embedded web server that reduces the design and implementation time of infrastructure, which drastically reduces the time cost of the project.

The access created is:
https://www.data-tigers.com/

Once there, you can find the current risk flooding map of Colombia where you can filter by place and date. It also contains historical data from 1998 to 2021, where you can deeply observe seasonal behaviour.

We kept just the information of *municipios* where weather stations are available.

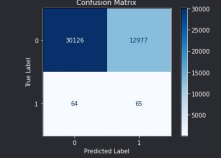
RESULTS

We determined the months where floods occur most often coincide with the months that are commonly the rainiest of the year, such as: April, May and November. We take the months 6, 7, 8, 9, 10 as low season time and the other months as high season time.

By breaking down the data by *departamento*, *year* and *month* we see some *departamentos* standing out by their high count of floodings across the historical data. Bogota D.C, Cundinamarca, Chocó, Risaralda and Antioquia make up the top 5 departamentos with more flooding reports historically.

The chosen model was an XGBoost model with Oversampling with the following results:

Geometric Mean Score: 0.5934



We found that flood events are more influenced by 'how fast' the precipitation was accumulated in a time frame, rather than the final total amount of rainfall in a day.

Remember to visit our website and protect your community! ASAP!

We have a mobile version.

You can also download and keep the information.