EARLY RISK ASSESSMENT OF COMMUNITIES PRONE TO FLOODING FROM PUBLICLY AVAILABLE GIS DATA IN COLOMBIA

PROJECT REPORT

TEAM 30: DATATIGERS | HTTPS://DATA-TIGERS.COM

2022

CONTENTS

..... 1

INTRODUCTION 4

PROJECT DEFINITION 5

OBJECTIVE 5

SCOPE 5

TECHNICAL FRAMEWORK 6

DATASETS 6

Overview: 6

Datasets Description: 7

DATA WRANGLING AND CLEANING 9

Format: 9

Relevance: 10

Data Augmentation: 11

EXPLORATORY DATA ANALYSIS 12

SOLUTION ARCHITECTURE 20

BACKEND 20

ML Model selection and tuning: 20

FRONTEND 26

Mockup: 26

Final Design: 27

Future Developments: 32

CONCLUSIONS AND TAKEAWAYS 33

HTTPS://DATA-TIGERS.COM

For primary user (landing page):

USER: protected
PASSWORD: protected

For dash app protection:

USER: protected
PASSWORD: protected

INTRODUCTION

As per data from the UNGRD's "Registro Único de Damnificados", for the period 2010 - 2011 the economic loss and damages due to natural disasters including floodings, landslides, and other climate associated events during rainy season in Colombia went up to $6.5 Billion USD (5.7% of GDP), leaving more that 2 million affected people.

> **"For the period 2010 - 2011 the economic loss went up to $6.5 billion USD (5.7% of GDP), leaving more that 2 million affected people."**

Rainy seasons usually involve floodings that cause infrastructure loss, agricultural assets damage, and food security issues. In Colombia, river flood hazard is classified as high based on modeled flood information. This means that potentially damaging and life-threatening river floods are expected to occur at least once in the next 10 years. An example of this is Mocoa flooding in 2017, where accordingly to Colombia Red Cross reports, more than 320 lives were lost and more than 15000 people were affected, completely or partially losing their houses, crops, livestock, or other assets.

Flash floods tend to be associated with many types of storms, all capable of producing excessive rainfall amounts over a particular area, so detection and prediction remains a challenge. In Colombia, the main tools used to detect and predict heavy rainfall associated with flash floods using traditional methods are based on the climate measuring stations network (rain, pressure, and temperature gauges) distributed across the country. Other information sources like satellite, lightning observing systems and radar are not extensively used in flooding forecasting.

Although UNGRD has implemented early warning systems to emit lifesaving alerts, this system is reactive and does not allow performing predictions of high-risk areas to facilitate the resources allocation and preventive actions; and usually, these warnings are not notified efficiently to the communities exposed to the risk.

Creating a model that would predict the high-risk areas based on publicly available data using Machine Learning models and creating a channel to communicate these early warnings easily and effectively to the impacted communities would enhance public entities resource allocation processes and would allow communities to perform preventive actions more effectively. This would improve the current Natural Disasters Response System capabilities, reducing the number of casualties and damages imparted on communities.

PROJECT DEFINITION

OUR OBJECTIVE IS SAVING LIVES

OBJECTIVE

Reducing the number of casualties and damages imparted on communities due to floodings in Colombia by creating a Machine Learning model that would predict the high-risk areas prone to flooding based on publicly available data and creating a channel to communicate these early warnings easily and effectively to the impacted communities.

WE WILL REACH THE TOP 17 MUNICIPIOS IN COLOMBIA WITH THE HIGHER COUNT OF FLOODING EVENTS AND AVAILABLE STUDY VARIABLES DATA

SCOPE

This project will be spatially framed to the municipios in Colombia with the higher count of flooding events and available public climate data (pressure, temperature, and precipitation) gathered from government official sources and flooding reports. 17 municipios were included in total. All other municipios in Colombia will be excluded from this project since no sufficient input data would be available to train the ML model to evaluate these cases.

WE WILL USE DATA SINCE JAN/2017

Temporarily, this project will be framed by the data available since Jan/2017 (5 years) and the moths with the higher count of flooding events across the country. This framing follows multiple purposes, but mainly is focused to reduce the imbalance between non-flooding events and flooding events to optimize the ML model development.

TECHNICAL FRAMEWORK

DATASETS

Overview:
The following list summarizes the data sources used to develop this project. Most of datasets were taken from official publicly available data sets published in official websites of government entities.

	NAME	SOURCE LINK	DESCRIPTION	YEARS AVAILABLE
1	UNGRD Emergencies	https://portal.gestiondelriesgo.gov.co/Paginas/Consolidado-Atencion-de-Emergencias.aspx	UNGRD: Emergencies databases per year	1998-2021
2	IDEAM Precipitation	https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Precipitaci-n/s54a-sgyg	IDEAM: Precipitation database	2003-2022
3	IDEAM Temperature	https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Datos-Hidrometeorol-gicos-Crudos-Red-de-Estaciones/sbwg-7ju4	IDEAM: Temperature database	2001-2022
4	HLS	https://hls.gsfc.nasa.gov/	Landsat and Sentinel-2 data: multispectral satellite measurement	
5	GeoJSON Col	https://www.colombiaenmapas.gov.co/	Colombia GeoJSON by department	2022
6	IDEAM Atmospheric Pressure	https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Presi-n-Atmosf-rica/62tk-nxj5	IDEAM: Atmospheric Pressure database	2001-2022
7	GeoJSON Col (By municipios)	https://www.colombiaenmapas.gov.co/	Colombia GeoJSON by municipio	2022
8	DIVIPOLA	https://www.datos.gov.co/widgets/gdxc-w37w	Standardized IDs of all municipalities in Colombia	2022
9	STRM - Altitude	https://developers.google.com/earth-engine/datasets/catalog/USGS_SRTMGL1_003	Digital elevation models	

Table 1. Datasets overview.

Datasets Description:

- UNGRD Emergencies

UNGRD Emergencies is a dataset created by the "Unidad Nacional de Gestion del Riesgo y Desastres" (UNGRD), this is the official Colombia government entity in charge of keeping record of all natural disaster events, implementing prevention and mitigation strategies, and coordinating immediate response in natural disasters events.

The dataset contains 52511 records of natural disaster events occurred across the country, from Jan/1999 to Dec/2021. It provides details on date of the event, location by departamento and municipio, type of event, affectations to community and supplies allocated to calamity response.

- IDEAM Precipitation

IDEAM Precipitation is a dataset created, maintained, and updated by the "Instituto de Hidrologia, Meteorología y Estudios Ambientales" (IDEAM), this is the official Colombia government entity in charge of measuring, monitoring, and forecasting climate variables like precipitation.

The dataset contains 168 million records of precipitation measurements distributed across 803 monitoring stations sensors located across the country, the information is updated with a delay of one calendar day back, containing data since January 2003 with a variable periodicity between 5 to 10 minutes for each sensor. The dataset provides details on measurement station id, station location by departamento, municipio and coordinates, date and hour of measurement, and measured value.

- IDEAM Temperature

IDEAM Temperature is a dataset created, maintained, and updated by the "Instituto de Hidrologia, Meteorología y Estudios Ambientales" (IDEAM). The dataset contains 66.2 million records of temperature measurements distributed across 559 monitoring stations located across the country, the information is updated with a delay of one calendar day back, containing data since January 2001 with a periodicity of 10 minutes for each sensor. It provides details on measurement station id, station location by departamento, municipio and coordinates, date and hour of measurement, and measured value.

- IDEAM Atmospheric Pressure

IDEAM Atmospheric Pressure is a dataset created, maintained, and updated by the "Instituto de Hidrologia, Meteorología y Estudios Ambientales" (IDEAM). The dataset contains 17 million records of atmospheric pressure measurements distributed across 358 monitoring stations located across the country, the information is updated with a delay of one calendar day back, containing data since January 2001 with a periodicity of 10 minutes for each sensor. It provides details on measurement station id, station location by departamento, municipio and coordinates, date and hour of measurement, and measured value.

- GeoJSON Col By departamento and GeoJSON Col By municipios

Plain file in open standard JSON format (JavaScript Object Notation) with the applied geometry based on the polygons determined by the GPS points that delimit the Colombian political division by departamento and by municipios respectively. It is important to mention that the identifier of each polygon corresponds to the one used by different public entities (DIVIPOLA), allowing generic information to be linked with georeferenced data.

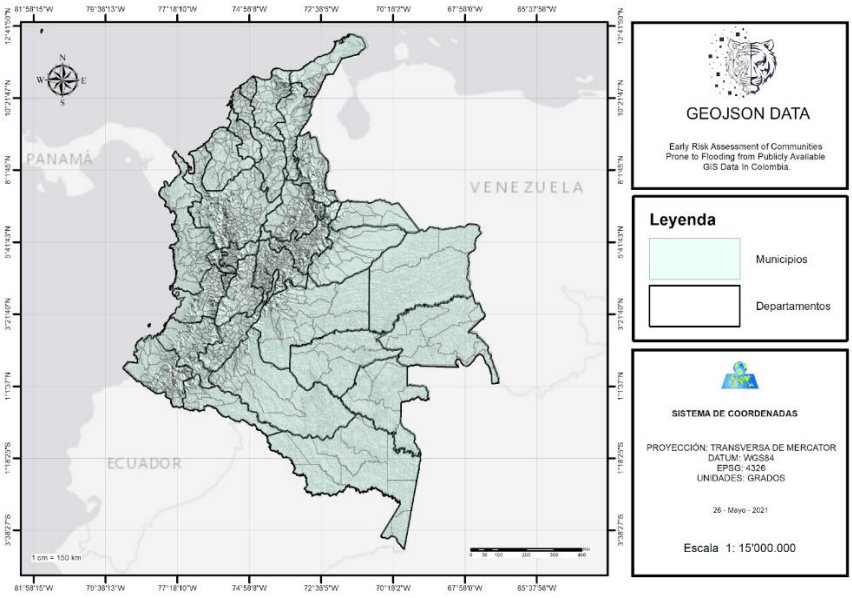

Figure 2. GeoJSON Colombia

- DIVIPOLA

DIVIPOLA is a dataset created by the "Departamento Administrativo Nacional de Estadística" (DANE), this is the official Colombia government entity in charge of measuring, monitoring, processing, analyzing, and diffusing the official statistics metrics of Colombia.

This dataset contains a standardized ID nomenclature designed by DANE for the identification of Territorial Entities, called Divipola. DANE defines Divipola[1] as a national standard that codifies and lists the territorial entities: departamento, municipalities, departmental townships, as well as population centers, police stations, hamlets, and municipal townships in rural areas.

DATA WRANGLING AND CLEANING

Format:

The UNGRD Emergencies dataset is published by UNGRD in its official website as a group of excel files, each one containing the reported events for every year since 1998. A total of 24 excel files can be found on the website at the moment. This means that before processing the data contained in these files, they need to be merged in a single dataset with a consistent format.

24

EXCEL REPORTS WERE PROCESED AND STANDARIZED TO UNIFY THE UNGRD EMERGENCIES DATASET UNDER THE SAME FORMAT

Excel files reports prior to 2019 are built following a different structure that those beyond that year. The file extension, number of columns and the quality of the data is different; therefore, several technical activities of unification were carried out with the objective of obtaining the appropriate format for all the data.

All 24 excel files were merged in a single dataset solving file extension issues, security risks issues originated by the automatic execution of unidentified macros, password protected files, and non-standardized entries.

A DATA PIPELINE WAS BUILT TO ACQUIRE AND PROCESS IDEAM DATA ON A DAILY BASIS

Once acquired the most recent published data from the official website of the UNGRD, the departamentos and municipios names were standardized based on information published by the DANE. Once the standardization process was covered, the Divipola ID number of each municipio was appended to the UNGRD Emergencies dataset accordingly to the DIVIPOLA dataset. Later, all columns containing similar information were unified based on the type of data and its concept or representation. Finally, data wrangling and data cleaning techniques were executed to the unified data, achieving a better data quality and a consistent format.

[1] https://www.dane.gov.co/files/investigaciones/divipola/divipola2007.pdf

On the other hand, the IDEAM precipitation, IDEAM Temperature and IDEAM Atmospheric Pressure datasets were acquired through a data acquisition and data processing pipeline that calls the most recent data from the official IDEAM website on a daily base. The Socrata API was used to build this data pipeline since is the only tool enabled by the publishing entity to perform this task[2].

Relevance:

As noted from the datasets description, the climate monitoring network does not have the same amount of precipitation, temperature and atmospheric pressure gauges distributed across the country. Figures 3, 4 and 5 shows the gauges distribution for each measured variable across the country versus the municipios with flooding reports, the deeper the dark blue the higher the flooding events count.

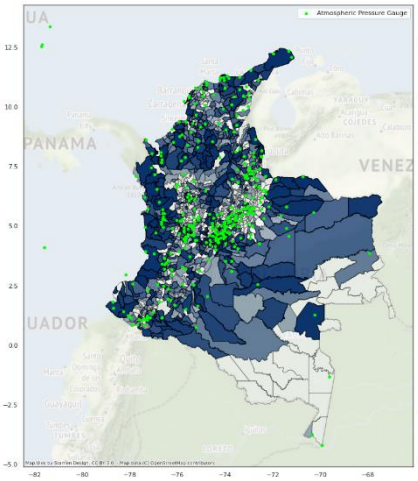

Figure 3. Atmospheric pressure gauge network (green dots) vs flooding reports by municipio (dark blue).

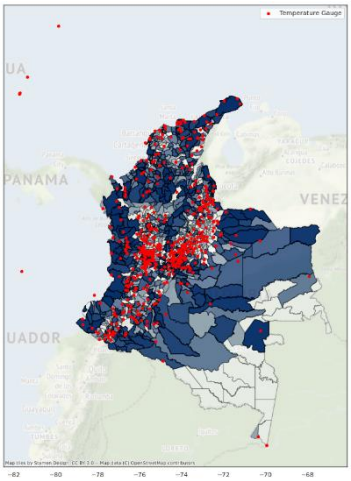

Figure 4. Temperature gauge network (red dots) vs flooding reports by municipio (dark blue).

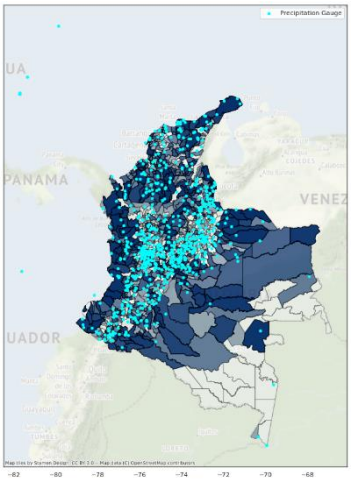

Figure 5. Precipitation gauge network (cyan dots) vs flooding reports by municipio (dark blue).

Since not all measurement networks have the same size, it implies that not all municipios in the country will have available measured data for all three study variables. This represents a big constraint for the development of a flooding forecasting ML model covering all municipios in the country. Given this restriction, and with no more official data sources available, this project scope was spatially framed to the municipios across the country with flooding reports in the UNGRD Emergencies dataset and available data for all three study variables: precipitation, temperature, and atmospheric pressure. A total of 248 out of 1142 municipios were included in this project given the previous criteria.

[2] https://dev.socrata.com/foundry/www.datos.gov.co/s54a-sgyg

Additionally, to reduce the imbalance between non-flooding days and flooding days, and increasing the representativity of data to the current environmental conditions, all datasets were temporarily framed to the values recorded beyond Jan/2017 (5 years of data).

Data Augmentation:

WE USED GOOGLE BIGQUERY TO EFICIENTLY TRANSFORM AND AUGMENTATE THE DATA

The project hypothesis revolves around finding a relation between the climate measurements and the flooding events in each municipio with available data. The IDEAM Precipitation dataset contains precipitation measurements reported every 5 or 10 minutes without a defined pattern in this sense. Giving the granularity level of this data and considering that the UNGRD Emergencies dataset do not provide details on the hour of occurrence of the flooding events, finding a relation between precipitation measurements and flooding events can be hard, resource expensive or even impossible.

To solve this issue, the IDEAM Precipitation data was aggregated by station and day-of-observation to obtain the daily maximum, minimum, mean, and standard deviation per weather station. Additionally, based on the principle proposed by Qian Ke et al[3], where the severity of the precipitation is determined by the volume amount of water that is poured in a given period of time, the precipitation data was aggregated to calculate a cumulative precipitation volume. The data was accumulated on different temporal scales, namely every 30, 60, 120, 350, 710, 1080 and 1440 minutes. Then the daily maximum precipitation volume at each temporal scale was selected.

Considering the data aggregation methodology discussed previously, it was necessary to use the Google Cloud Platform's BigQuery service to call the data from IDEAM official website and transform the data. BigQuery service leveraged the acquisition and data transformation pipeline workflows obtaining calculated data in an expeditious way for daily rainfall measurements per sensor, and cumulative precipitation calculation for temporal scales of 10, 30, 60, 120, 120, 360, 720, 1080 and 1440 minutes.

[3] Qian Ke, Xin Tian, Jeremy Bricker, Zhan Tian, Guanghua Guan, Huayang Cai, Xinxing Huang, Honglong Yang, Junguo Liu, Urban pluvial flooding prediction by machine learning approaches – a case study of Shenzhen city, China, Advances in Water Resources, Volume 145, 2020, 103719, ISSN 0309-1708

EXPLORATORY DATA ANALYSIS

This project is constrained by the availability of official data included in climate datasets reported by IDEAM. Since the reported data has a delay of one day, the ML model and therefore the EDA of this project will be build based on the hypothesis that yesterday's climate data is related to today's flooding events.

Out of 18042 events recorded in the UNGRD Emergencies dataset since Jan-2017, 2708 are flooding events representing around 15% of total recorded emergencies, being by far the most common emergency event. Once the UNGRD Emergencies dataset is merged with IDEAM climate datasets, the number of flooding events with available data for all three variables is reduced to 635 records distributed across 145 municipios.

Figure 6 shows the distribution of the cumulative precipitation at time equal to 1440 min (one day) for the 635 records mentioned previously showing a right skewness distribution heavily concentrated in cumulative precipitation values below 50 mm. This suggest the presence of flooding events where the cumulative precipitation at time equal to 1440 min is zero, meaning that the dataset includes flooding events that where not caused by rain. A total of 113 events were excluded from the analysis and ML model since are not included within the scope of this project given the independency of those flooding events with the climate behavior variables. 522 flooding records are left after excluding the events previously mentioned.

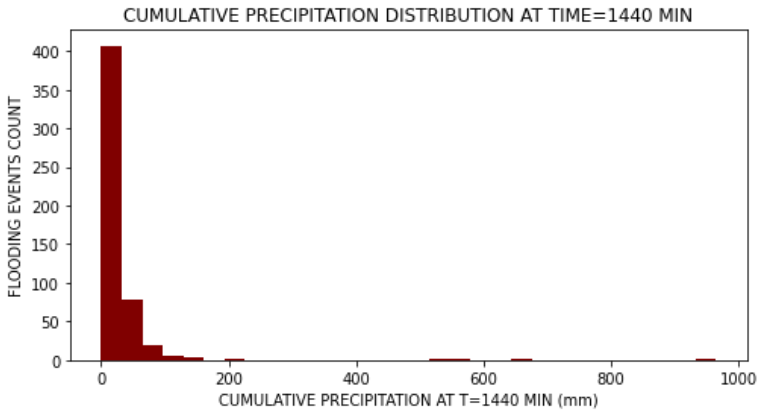

Figure 6. Cumulative precipitation distribution at time equal to 1440 min (one day).

As described in figure 7, just a few municipios account for around 50% of the flooding records in the merged dataset. Bogota, Pereira, Villavicencio, Santa Marta and Cali make up the top 5 municipios with

the highest count of flooding records. To maximize the amount of flooding days while minimizing the amount of non-flooding days, and therefore reducing the dataset imbalance to optimize the ML model, this project scope was framed to the top 17 municipios where the dataset imbalance reaches its optimal point. The top 17 municipios make up for the 50.96% of total flooding-days records included in the dataset while containing only 17.49% of non-flooding-days records.

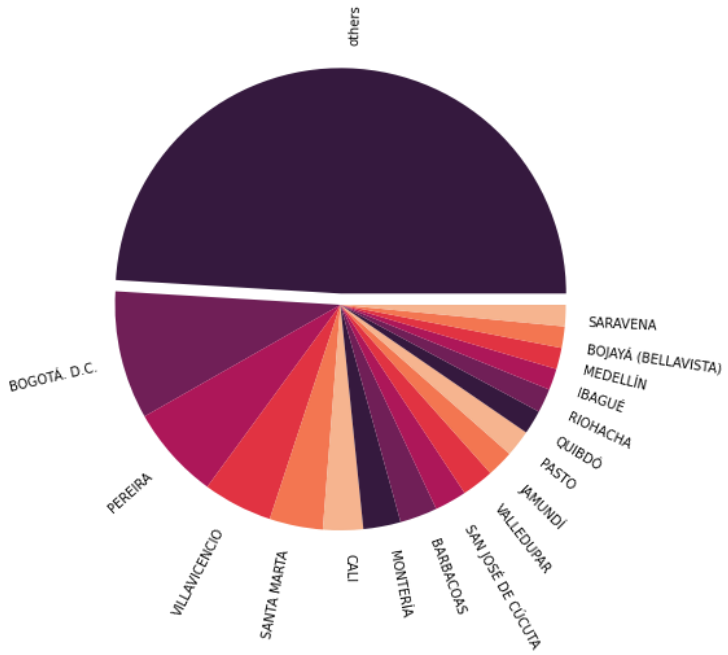

Figure 7. Flooding records count distribution by municipio.

By exploring the occurrence of flooding reports for the 17 selected municipios, it was found that the historically the months with the lower count of flooding reports are January, February, September, and December. Likewise, the months with the higher count of flooding reports are May, June, October, and November, following a bimodal tendency. The relation between month and flooding event occurrence is confirmed by running a chi-square test obtaining a P-value of 1.1079e-14, discarding the null hypothesis and suggesting a relation between the two variables.

The dispersion of the flooding events count aggregated by month can be easily explored in figure 9. As expected, the months with the lower count of flooding reports are also the ones with the lower dispersion of data. Likewise, the months with the higher count of flooding reports are also the ones with the higher dispersion of data. Bogota D.C., Pereira, Villavicencio, Santa Marta and Cali, the top 5 municipios with the highest count of flooding reports, are the municipios accounted in a high degree for the dispersion of the Top 17 municipios dataset.

Although the median for all months tends to be around 2 flooding events per month, the data for May, June, October, and November are highly skewed, indicating the presence of municipios with a non-typical flooding events count adding a high dispersion to the values in the months listed previously. The top 5 municipios are the ones accounted for this behavior.

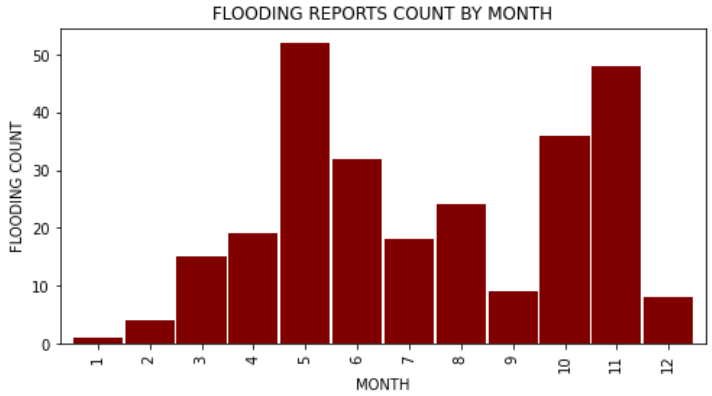

Figure 8. Flooding reports count by month.

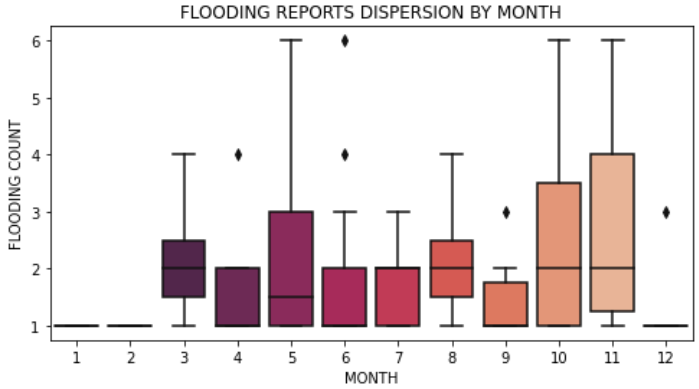

Figure 9. Flooding reports count dispersion by month.

The same bimodal tendency can be noted in figure 10 by breaking down the data by municipio, year and month. It can also be noted that the data has a significant reduction in reported flooding events for the 2020 period, displaying a low count of flooding reports. This might suggest that the UNGRD Emergencies dataset includes a bias originated in lack of reported events for municipios across the country due to the challenges that COVID-19 pandemic represented back in the time.

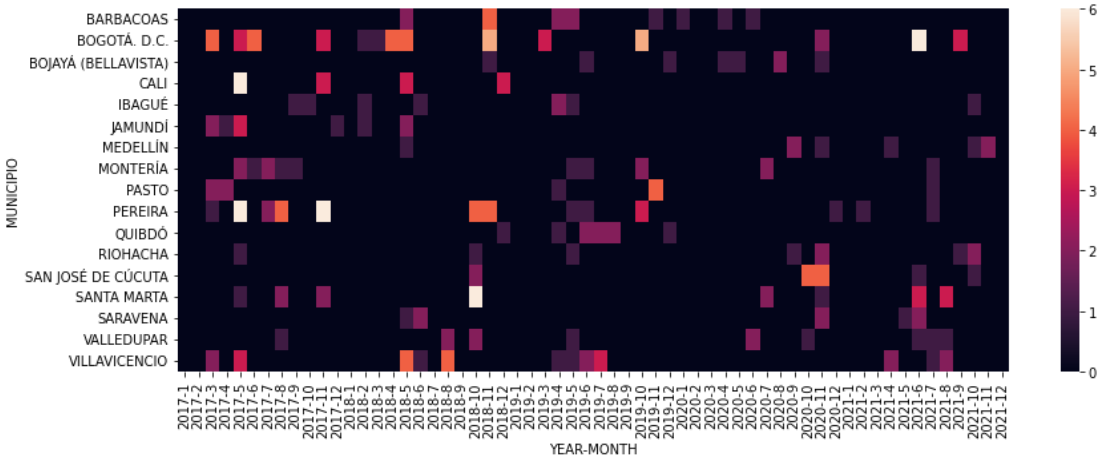

Figure 10. Heatmap of flooding reports count by municipio, by year and month.

This bias becomes more evident by looking back in more detail figures 3, 4 and 5, where the municipios with flooding reports are colored with blue. All municipios in the south-wests region of Colombia, within the Amazonian hydrographic basin, are significantly underrepresented when compared to other regions of the country. Is well known that due to the environmental conditions particular to the Amazonian rainforest, these municipios are highly prone to flash floodings. This suggests a representation bias of the UNGRD Emergencies dataset. Additionally, there is a large cluster of measuring stations around the Andean area of the country.

The density kernel plot shown on figure 11 displays the calculated density of stations across the country around the point-type entities represented by the IDEAM stations for the year 2020. It can easily be concluded that measurement stations are heavily concentrated across the Andean region in the country, making climate data not available for municipios far from the high-density areas.

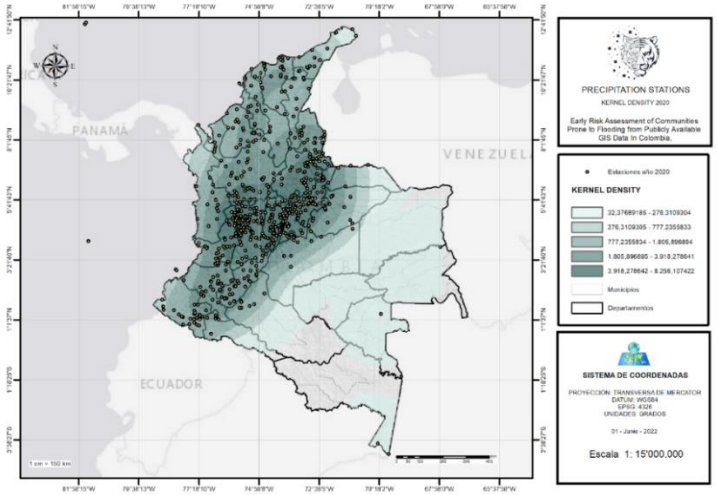

Figure 11. Kernel density map for IDEAM stations.

The cumulative precipitation by month since 2017 for the selected 17 municipios is plotted in figure 12. At a first glance, it can be noted that the bi-modal tendency for flooding events described in figure 8 is not aligned with the tendency of cumulative precipitation, where March is the month with the highest cumulate precipitation value. This might suggest that the occurrence of flooding events is not related to the total cumulative precipitation. However, if the data is disaggregated by month and year, as shown in figure 13, it can be noted that Q4-2018, Q1-2019 and Q2-2019 were atypical quarters along historical data where precipitation values were way higher than usual, affecting the aggregated data behavior. The independence of both variables is discarded by performing a chi-square test between Flooding Events Occurrence and Cumulative Precipitation at different times (10, 30, 60, 120, 360, 720 and 1440 min) obtaining P-values in the range from 3.6244e-221 up to 2.2551e-60, suggesting a relation between the cumulative precipitation variables and the occurrence of floodings.

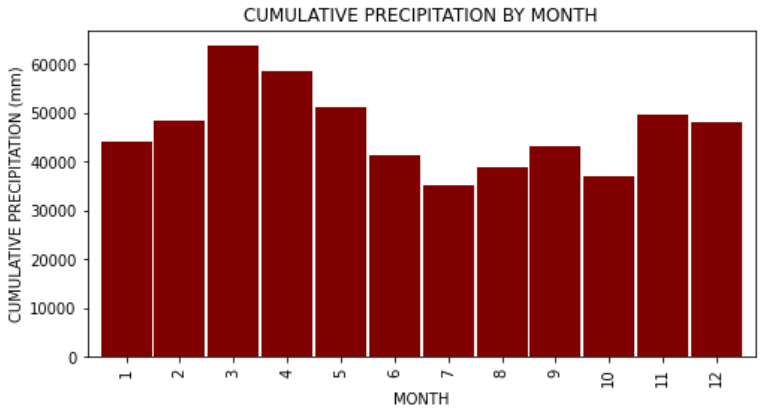

Figure 12. Cumulative precipitation by month.

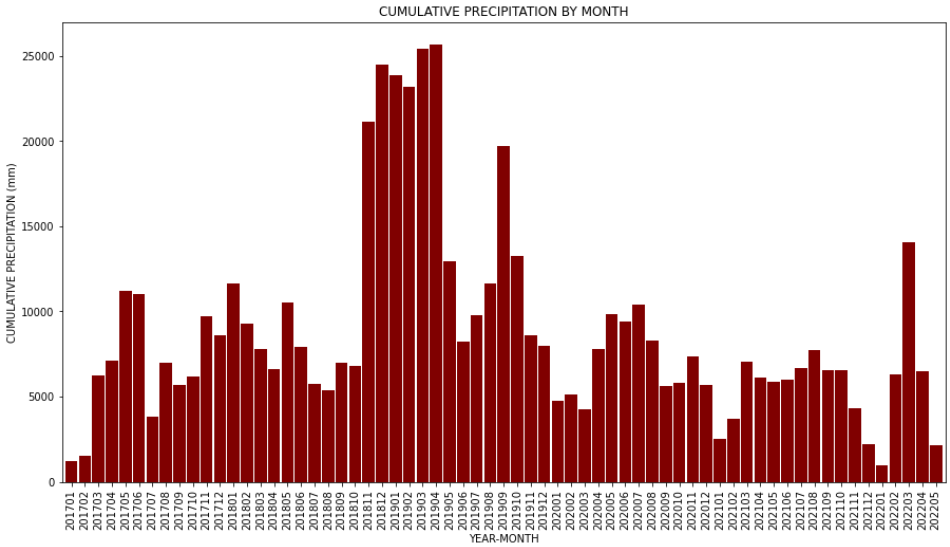

Figure 13. Cumulative precipitation by year and month.

Total cumulative precipitation cannot explain the occurrence of flooding events by itself, that is why is needed to add an additional dimension to the cumulative data by evaluating the hole behavior of cumulative curves versus flooding occurrence.

For the particular case of Colombia, all months have a count of flooding, but some have a much higher count that others, therefore the months were divided in three groups accordingly to figure 8 distribution. The high flooding count months (May, June, October, and November), the low flooding count months (January, February, September, and December) and the months that do not fall on any of the previous categories were grouped as middle flooding count months. The behavior of the cumulative precipitation curves is detailed on figure 14. As noted, the higher the cumulative precipitation curve is at early times, the higher the flooding count; meaning the faster a volume is released by rains, the higher the chances of floodings to occur.

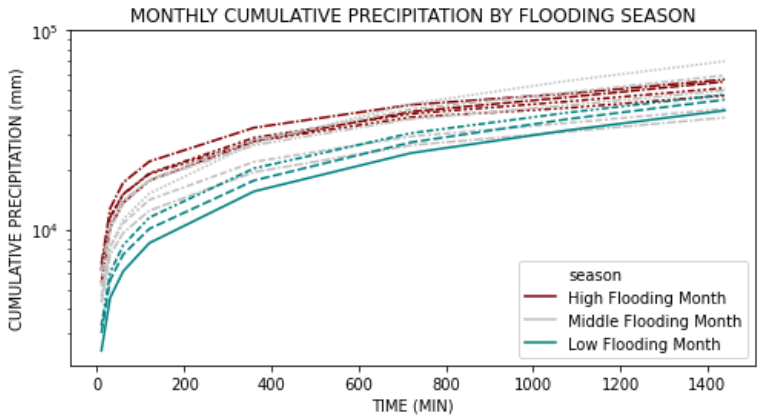

Figure 14. Cumulative precipitation curves for high, middle and low flooding count months.

The relation of the temperature values and atmospheric pressure values are well known to be highly related to the precipitations. These climate variables and their relation are extensively used in conventional models to perform rain forecasting. Both temperature and atmospheric pressure follow an inverted behavior when comparted to precipitation, the higher the temperature and atmospheric pressure, the lower the precipitation. These datasets will support the relations provided by the IDEAM Precipitation dataset when related to flooding occurrence, improving the ML model performance.

On the other hand, the volatility of measurements on both variables increases considerably when disaggregating by measurements recorded on flooding days and measurements recorded on non-flooding days. High volatilities tend to be associated with high flooding count months.

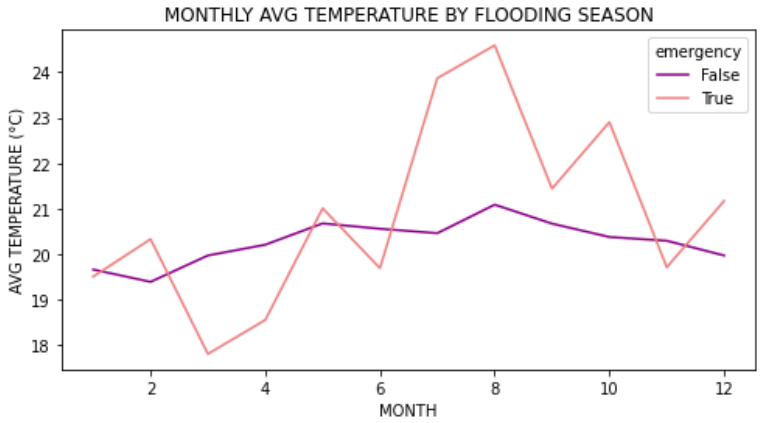

Figure 15. Average temperature curve by month for flooding (True) and non-flooding (False) days.

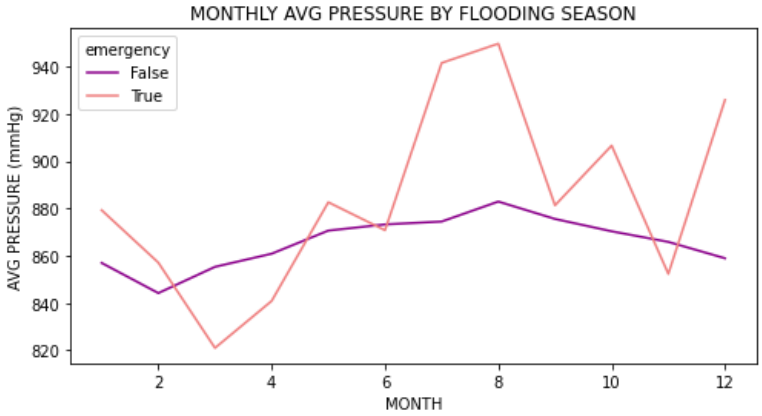

Figure 16. Average atmospheric pressure curve by month for flooding (True) and non-flooding (False) days.

The strip plot below describes the atmospheric pressure measurement recorded in the period from Jan/2017 up to Dic/2021 for Bogota D.C. measurement stations. Red dots represent those measurements corresponding to days where floodings were reported, while the gray dots represent the historical measurements of the same measurement stations across the same period. Thanks to this view we can identify a clear pattern of atmospheric pressure vs flooding events that could lead to a positive correlation between the variables. If this hypothesis is confirmed, we could use atmospheric pressure records to leverage the flooding risk prediction workflow.

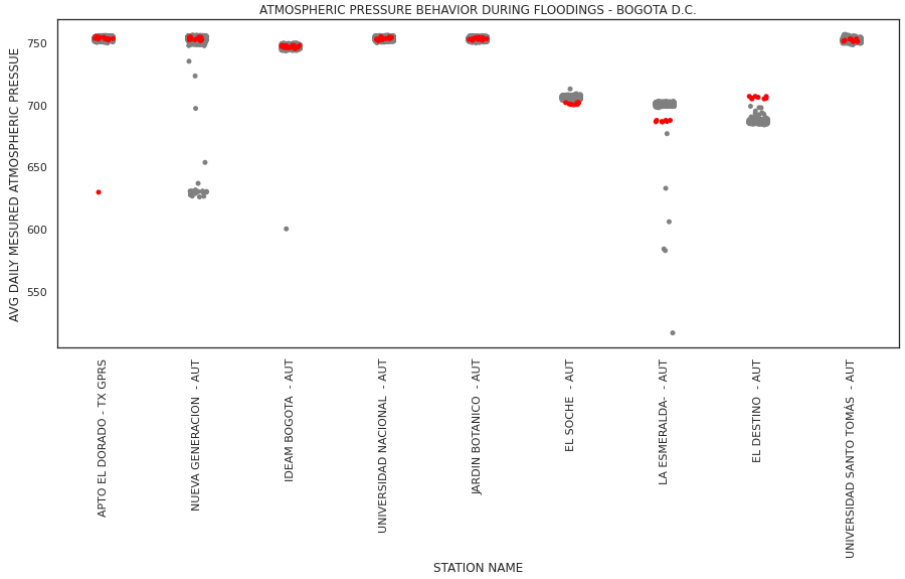

Figure 16. Average daily atmospheric pressure by measurement station in Bogota D.C., grouped by flooding (red) and non-flooding (gray) days.

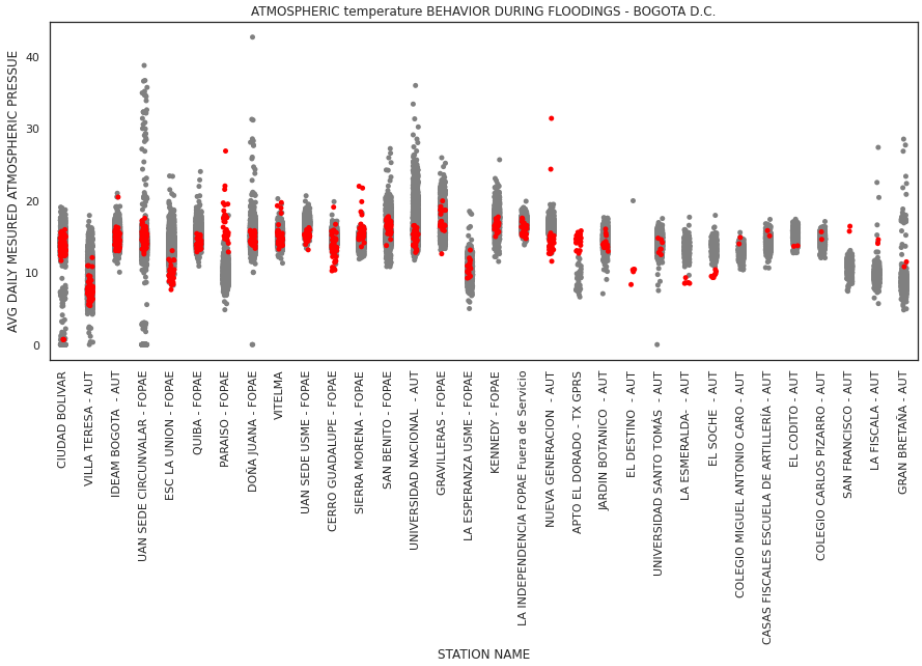

Figure 17. Average daily atmospheric temperature by measurement station in Bogota D.C., grouped by flooding (red) and non-flooding (gray) days.

The strip plot above describes the temperature measurements during recorded flooding events for Bogota D.C. (red) and the historical data of the same measurement stations across the city (gray). Thanks to this view we can identify a clear pattern of temperature vs flooding events that could lead to a positive correlation between the variables. If this hypothesis is confirmed, we could use temperature records to leverage the flooding risk prediction workflow in later stages of this project.

SOLUTION ARCHITECTURE

BACKEND

ML Model selection and tuning:

The various datasets preprocessed beforehand during the EDA were combined in order to create the final base table to serve the model. That is, the IDEAM Precipitation, IDEAM Pressure, and IDEAM Temperature were combined using the codigo_estacion, nombre_estacion, departamento, municipio, anio_observacion, doy_observacion, day_observacion, mes_observacion, zona_hidrografica, and DIVIPOLA, forming a unique dataset with all the predictors for the model. The UNGRD Emergencies dataset (which contained the target variable for the model), however, was combined with the predictors with a one-day delay with respect to the measurement date. As previously discussed, the model was created to return the probability of an emergency happening today based on the measurements taken yesterday.

From the resulting dataset, it was noticed a large number of floods that were not related to precipitation during the previous days. They are likely occurring:

- When a nearby river overflows due to high precipitation upstream
- When a sewer system fails
- Lack of accurate precipitation data from the government entity when the emergency occurred

Consequently, those events were excluded as previously discussed, feeding the model only with those flood events in which there were rain measurements the day before. That is: the model was designed to predict the probability of rain-related floods.

The correlation plot between the resulting predictors is shown in the graph below. Most of them are self evident and expected. For example, there is a large relationship between the min, max, avg, and std of each individual measurement in the dataset.

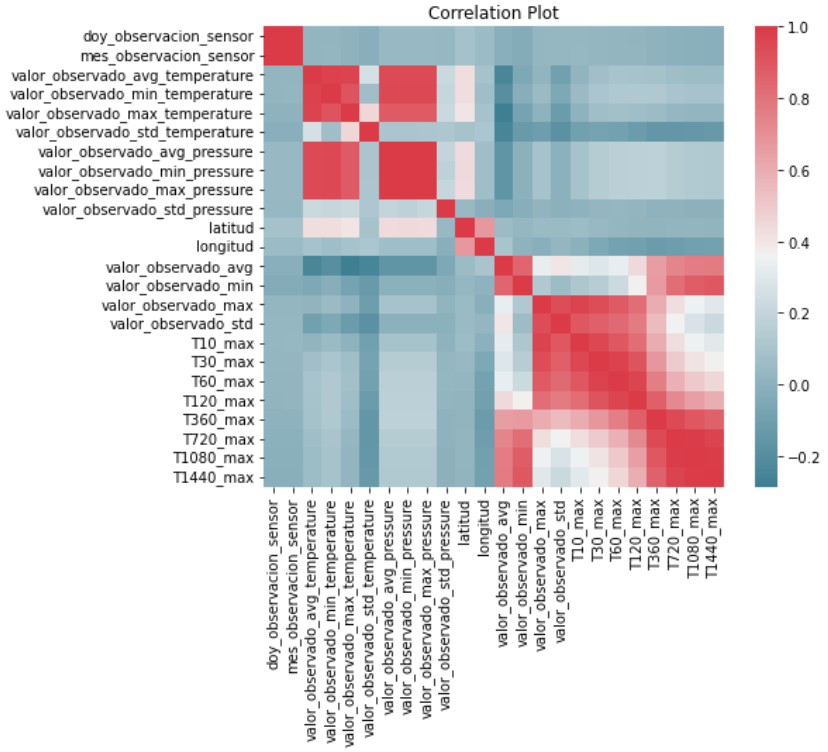

Figure 18. Correlation plot between resulting predictors.

Additionally, a Chi-square test was ran on the categorical variables (zona hidrográfica and DIVIPOLA) against the target variable. The results are as follows:

- P-value for Zona Hidrográfica: 0.000516071732151843
- P-value for DIVIPOLA: 9.815771674488654e-06

The interpretation of the P-values above is that there is a significant relationship between the categorical variables and the occurrence or not of emergencies.

The split between training and test sets was not random or stratified, as is the norm in most cases. Instead, it was decided to use historical data from 2017 to 2020 for training, and 2021 data for testing. This approach was implemented to avoid data leakage and to simulate the behavior the model will face in real life, where the model will be trained with historical data and used to predict the current year.

Additionally, it was found that floods are a dangerous but rare occurrence and this is evident in the large imbalance between the classes (99.6% of the time there is no emergency). As previously discussed, the project was framed to 17 municipios, which were those in which a large number of flooding emergencies had occurred between 2017 and 2021. These municipios are: Bogotá. D.C., Pereira, Villavicencio, Santa Marta, Cali, Montería, Barbacoas, San José de

Cúcuta, valledupar, Pasto, Jamundí, Riohacha, Quibdó, Ibagué, Medellín, Saravena, and Bojayá.

This approach slightly decreased the imbalance, but the positive class still was present in only 1.2% of instances of the dataset. Thus, a couple of additional measures were implemented in order to reduce this problem. The first was to use a sampling method like random oversampling and/or undersampling to balance out the classes in the training set. The second was to use the Geometric Mean Score as the target metric for the model tuning, as suggested in Machine Learning forums online[4]. This score is designed to maximize the accuracy on each of the classes while keeping these accuracies balanced, and is defined as

$$\left(\prod_{i=1}^{n} Sesitivity_i\right)^{1/n}$$

where *i* represents each class (2 in our case).

It's also important to mention that the reason to use Random Oversampling instead of SMOTE to balance out the classes: SMOTE creates fake emergencies with values close to the real ones, and given that we have one-hot-encoded our categorical variables, SMOTE would assign values different than {0, 1} but somewhere in between, resulting in invalid ranges for these variables.

Various models were trained, including Random Forests, XGBoost and LightGBM and tuned their hyperparameters using a grid search. A summary table of their results is presented below:

Model Description	Confusion Matrix	Geometric Mean Score
XGBoost with undersampling of major class	2934, 1997 12, 28	0.6453
LightGBM with undersampling of major class	2842, 2089 16, 24	0.588
LightGBM with oversampling of minor class	1020, 3911 3, 37	0.4374
Random Forest with undersampling of major class	2478, 2453 12, 28	0.5931
Random Forest with oversampling of minor class	4931, 0 40, 0	0

Table 2. Trained ML models summary.

Additionally, as many hyperparameters as computationally possible were tuned for each model in a grid (sometimes randomized) search. The full list of parameters and their respective ranges analyzed are included in the table below.

[4] https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/

Model	Hyperparameter	Range of values
Random Forest	n_estimators	50, 100, 200, 500
Random Forest	max_features	sqrt, log2
Random Forest	max_depth	1, 2, 3, 4
Random Forest	criterion	gini
Random Forest	bootstrap	True, False
XGBoost	min_child_weight	sp_randInt(1, 10)
XGBoost	gamma	sp_randFloat(0.5, 5)
XGBoost	subsample	sp_randFloat(0.5, 1.0)
XGBoost	colsample_bytree	sp_randFloat(0.5, 1.0)
XGBoost	max_depth	sp_randInt(3, 5)
XGBoost	scale_pos_weight	np.linspace(100, 400, 4)
LightGBM	num_leaves	np.linspace(2, 10, 5)
LightGBM	max_depth	np.linspace(1, 10, 5)
LightGBM	is_unbalance	True, False

Table 3. Trained ML models parameters.

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

Geometric Mean Score: 0.6343. Notice that this score is slightly lower than the one obtained using Undersampling (see table above). However, there is a benefit to using this model on the entire dataset, as we'll see below.

ROC Curve and AUC: 0.6729

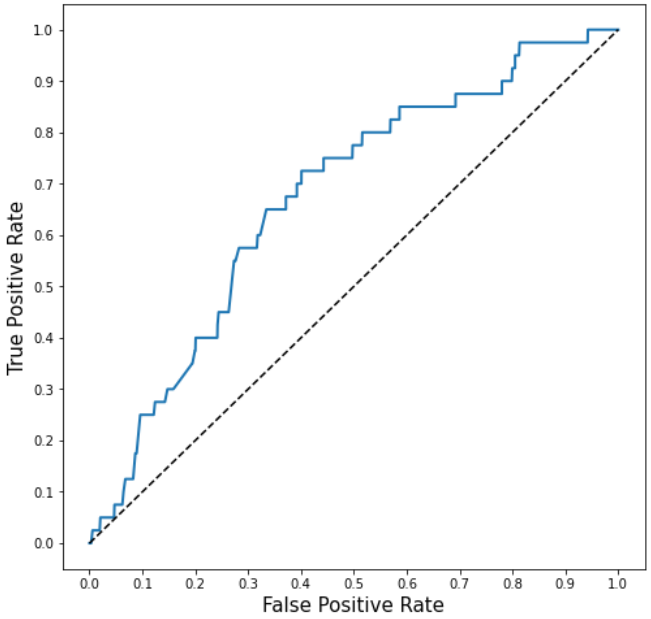

Figure 19. ROC curve for selected ML model.

Confusion matrix. Notice that 30 out of 40 emergencies are correctly classified, but 46% of predicted emergencies are false positives. These could be due to very rainy days in high probability areas, and reporting a high probability of an emergency can still be useful for its inhabitants to be prepared and take preventive actions.

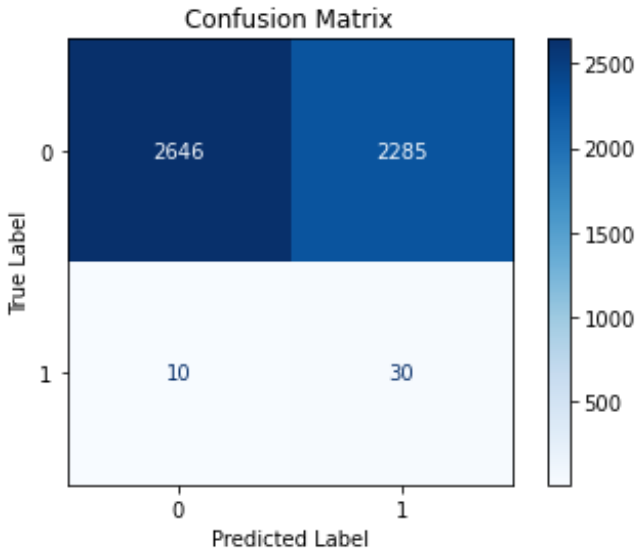

Figure 20. Confusion matrix for selected ML model.

Running the same prior measurements with the winning model but this time for all municipios and all emergencies of Colombia is important given that the model will help us make predictions everywhere within the country. Remember that the original dataset was filtered to only 17 municipios with high probability of having an emergency in order to reduce the data imbalance, and we kept only emergencies in which there was evidence of rain. The results are shown below:

Geometric Mean Score: 0.5934. Notice that this score remains relatively high even when using it on the entire dataset. This is desirable, as we want to keep the highest sensitivity in both classes when doing the predictions using real data.

ROC Curve and AUC: 0.6313

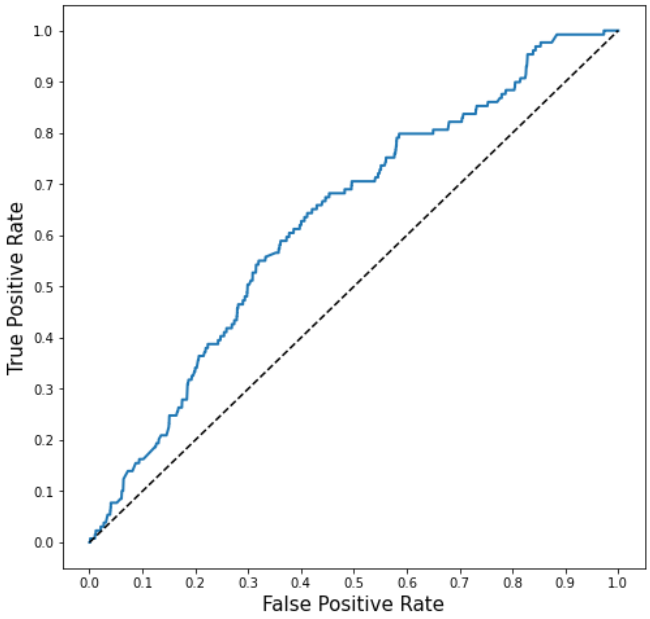

Figure 21. ROC curve for selected ML model – entire dataset.

Confusion Matrix. Notice that, given that we have the full 2021's dataset of emergencies, the specificity is relatively high, which is something we want. Additionally, out of the 64 emergencies incorrectly labeled by the model (false negatives), we can assume that many of them are related with variables outside of the temperature, pressure, precipitations, etc, as mentioned above. Finally, we can conclude that the results are far from ideal, but given the difficulty of this prediction, we can say that the chosen model does a good job, and can be improved in the future with more accurate data and new variables.

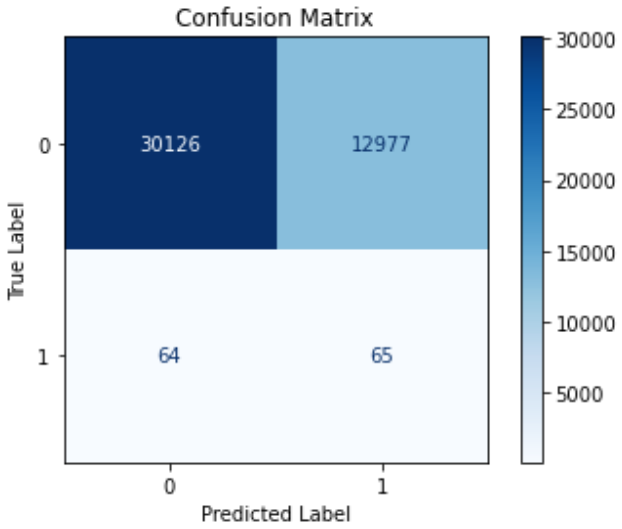

Figure 22. Confusion matrix for selected ML model – entire dataset.

FRONTEND

Mockup:

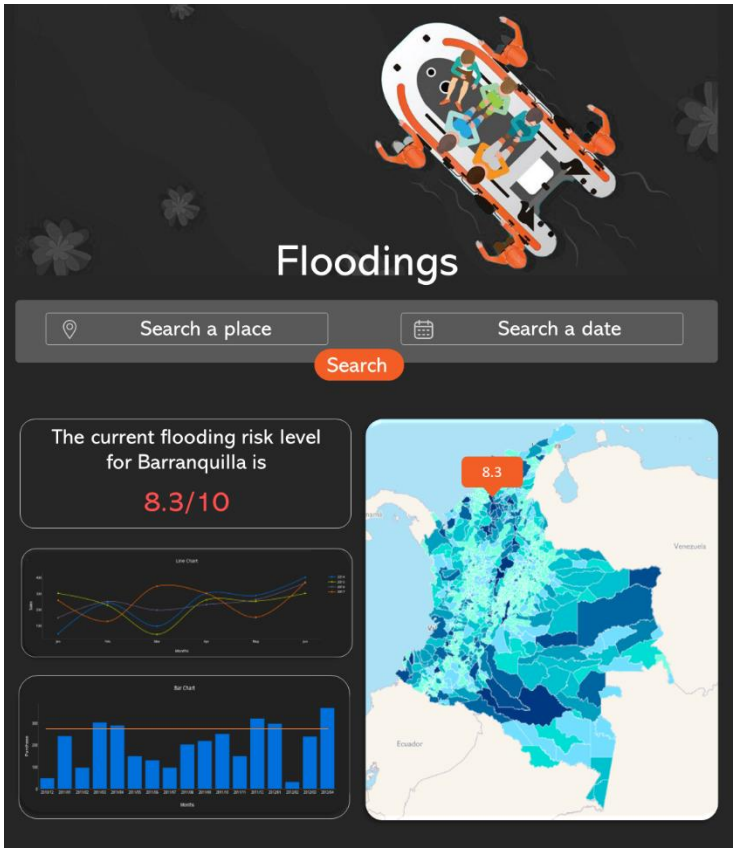

Figure 23. FrontEnd MockUp.

The initial approach used to design the model frontend was developed with a simple idea in mind: “keep it simple”. This gains relevance especially when considering that the final user that is going to be consuming the model outputs have a wide profile description and a multicultural background.

The required inputs by users will be limited to two searching criteria: location and date. By clicking the search button, users will be able to see the current flooding risk for the selected location; also will find a time series plotting the historic flooding risk for the last 6 months and a bar chart displaying the historic number of floodings for the selected location and the selected month.

The map will display a heatmap indicating the flooding risk across all municipios in Colombia. When a location search criteria is used, the map will zoom in into the desired location and will display any relevant data using a pop up marker in the map.

This project will use the DASH platform for the FrontEnd and database will be supported by PostgreSQL and Google Big Query, our team came to this conclusion after conducting a detailed study of the skills and abilities of the tools, thus:

- DASH:

The Dash platform empowers data science teams to focus on the data and models, while producing and sharing enterprise-ready analytic apps. Dash is a python framework created by plotly for creating interactive web applications written on the top of Flask, Plotly.js and React.js and has an embedded web server that reduces the design and implementation time of infrastructure in the development environment, which drastically reduces the time cost of the project.

- PostgreSQL:

PostgreSQL is an open source object-relational database system that uses and extends the SQL language combined with many features that safely store and scale the most complicated data workloads, our project benefits from PostgreSQL because it supports both SQL (relational) and JSON (non-relational) queries.

- GCP - Big Query:

BigQuery is a Google Cloud Platform (GCP) service with serverless architecture that uses SQL queries to answer the biggest questions with zero infrastructure management. BigQuery supports our project because its engine makes it possible to query terabytes of datasets in seconds and petabytes in minutes. The BigQuery ML documentation helps data scientists to discover, implement, and manage data tools to inform critical business decisions.

Final Design:

The entire FrontEnd was fully developed and designed in Dash version 2.5.1 with its respective components (all updated as of July 7, 2022). However, all requirements are recorded in the corresponding file: requirements.txt attached to this document.

- Responsive Design

The development was based on a responsive design that adapts the application design seen on the web page (www.data-tigers.com) to the client's viewing environment through the use of techniques such as fluid proportion-based grids and flexible images.

For this purpose, the dash-bootstrap-components library was used and properly configured because it's a library of Bootstrap components for Plotly Dash, and makes it easier to build consistently styled apps with complex and responsive layouts.

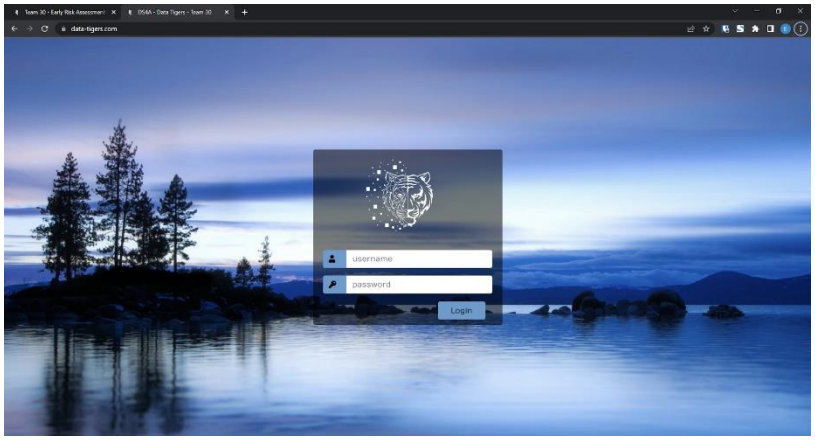

Figure 24. Landing page on https://www.data-tigers.com.

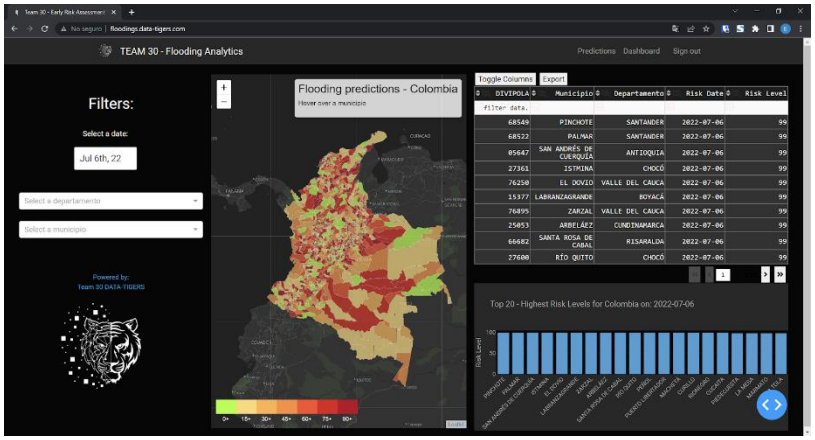

Figure 25. App in operation from web browser on PC.

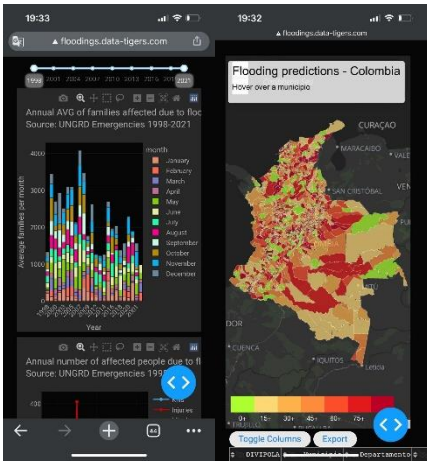

Figure 26. App in operation from web browser on PC.

Basically, the FrontEnd consults the emergency Datasets, DIVIPOLA, GeoJson by municipality described extensively in this document and at runtime smaller GEOJSON files are generated by date and department, with the necessary information to be able to plot the map of Colombia at the municipality level, then, with this done, the risk levels provided by the model are consulted from the BackEnd and updated daily to the application.

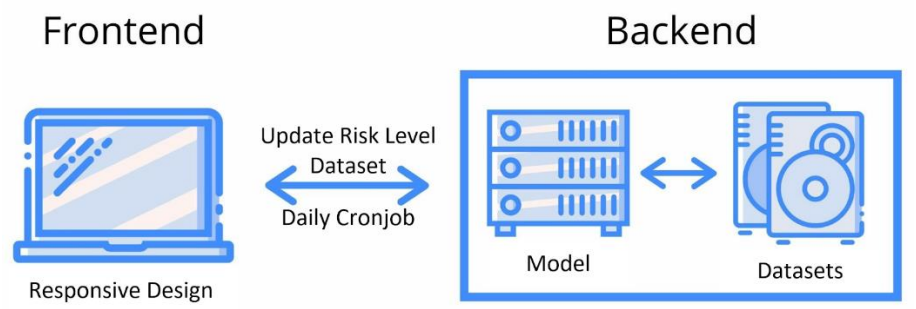

Figure 27. Frontend and Backend communication schema.

Then, for the operation of the application, the FrontEnd communicates with the BackEnd through the consultation of an intermediate dataset called risk_level.csv.

➢ Callbacks

Whenever an input like date, departamento, municipio, export button, or date slider are modified by the user, one function that the callback decorator wraps will get called automatically. Dash library provides this callback function with the new value of the input or group of inputs property as its argument, and Dash updates the property of the output or outputs component with whatever was returned by the function.

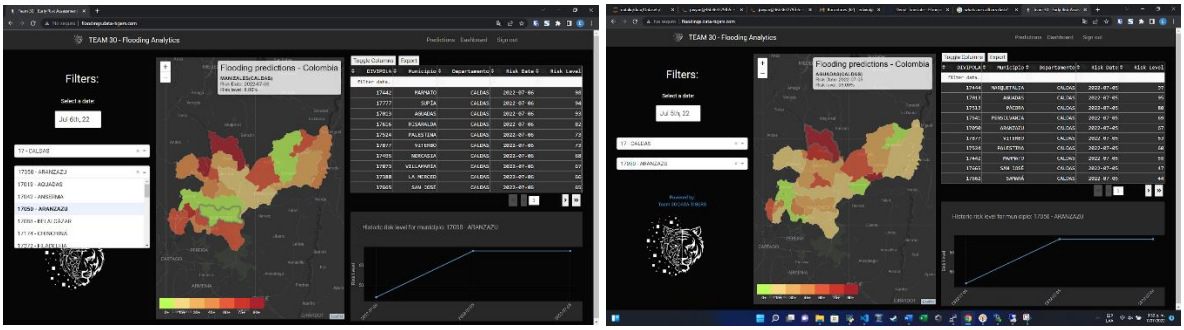

Figure 28. Callbacks in action: Date, Departamento, Municipio and Hover on Map

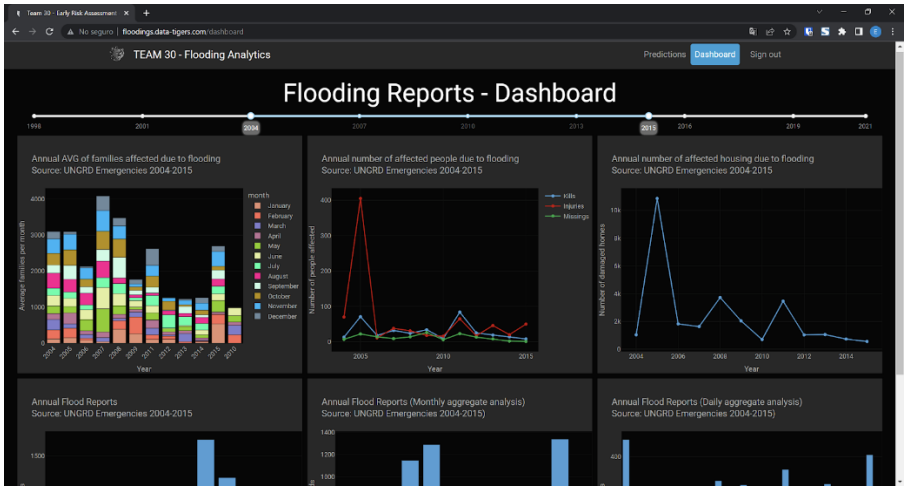

Figure 29. Callback for rangeSlider (Years on Dashboard)

A total of 4 callbacks were developed, one of then (update_map_table_graph) is a long callback (more complex)

```
#Callback for Navigationmenu
@app.callback(
    Output("app-content", "children"),
    [Input("url", "pathname")])
def render_page_content(pathname):
```

Figure 30. Definition of callback for Navigation menu

```
# Callback for hover over Map
@app.callback(
    Output("info_tool_tip", "children"),
    [Input("geojson", "hover_feature")])
def info_hover(feature):
```

Figure 31. Definition of callback for hover over Map

```
#Callback for date, list departamento and list municipio
@app.callback(
    Output('list-municipio', 'options'),
    Output('table-risk-level', 'data'),
    Output('graph-risk-level', 'figure'),
    Output('geojson', 'url'),
    [Input("my-date-picker-single", "date"),
     Input("list-departamento", "value"),
     Input("list-municipio", "value"),],
    running=[
        (Output("my-date-picker-single", "disabled"), True,
False),
        (Output("list-departamento", "disabled"), True, False),
        (Output("list-municipio", "disabled"), True, False),
    ],
    prevent_initial_call=True,
)
def
update_map_table_graph(selected_date,selected_departamento,select
ed_municipio):
```

Figure 32. Definition of callback for date, list departamento and list municipio

```
#Callback for rangeSlider (Dashboard)
@app.callback(
```

```
    Output('flooding_by_year', 'figure'),
    Output('flooding_by_month', 'figure'),
    Output('flooding_by_calendar_day', 'figure'),
    Output('affected-families', 'figure'),
    Output('affected-people', 'figure'),
    Output('affected-housing', 'figure'),
    [Input('slider-anios', 'value')])
def update_dashboard(value):
```

Figure 33. Definition of callback for rangeSlider (Years on Dashboard)

- Temporal and spatial complexity

The execution of the application implies the use of large computational resources, this is due to the constant intersection of large datasets and the join with GeoJson files. For this reason, 3 possibilities were analyzed to visualize the information in a geospatial way, all made in Spyder version 5.1.5 of Anaconda3 Environment

1. GeoJson to Dataframe (geopandas) joined with risk level (DataFrame): It's the worst option, the spatial and temporal complexity is very high, the RAM used in the tests was greater than 32GB and the time required per query or callback exceeded 12 seconds.
2. Filtered GeoJson to shape file (geopandas) only by deparments' geometry of displayed on screen and joined with risk level (DataFrame): The option reduces the spatial complexity (12GB in RAM approx) but the temporal complexity remains high: 8 seconds on average per data query.
3. Filtered GeoJson to geofile with minimun columns (Date and Risk Level per municipio) then open the new file and join with risk level dataset: The best option, the time and space complexity were reduced to the maximum, obtaining queries in less than half a second with RAM under 8GB, the cost of that: disk space.

- FrontEnd Security

Several configurations were made for the security of the application and the project was based on good practices of secure web development, such as the following:

SSL certificate enabled for the data-tigers.com domain (landing page and entire website)

Perimeter protection for the data-tigers.com domain and *.data-tigers.com top-level subdomains (the application doesn't use deepest level domains).

1. Two-phase user authentication, one for the web session and cookie generation specific to each visit and another for application execution protection, this second is an additional security measure enabled because the landing page is hosted on a different site where the application is running.
2. All other ports than 80 and 443 were blocked for the publication of the app, internally the firewall (Sophos Appliance) is responsible for protecting the perimeter and for external to internal communication the firewall makes NAT to ports 8050 (Dash) and 8888 (Jupyter hub) and thus these are not exposed to the Internet.
3. The IDS/IPS service monitors user behavior in the app and keeps a record of accesses, errors and connections made (audit logs).

- User authentication

As mentioned above, two valid login credentials are required to access the flood prediction and analysis application

For primary user (landing page):

USER: protected
PASSWORD: protected

For dash app protection:

USER: protected
PASSWORD: protected

Future Developments:

1. State-of-the-art user authentication: Dash has a service and specialized libraries for state-of-the-art user authentication such as Auth2.0, 2FA, passwordless etc., however, they are only enabled in DASH enterprise license for companies with high costs that can be analyzed for future versions.
2. Map service: The map service associated with the project is from Stadia Maps with free account, this is adequate for the project, but the layers enabled for the free versions of the service are limited and the combination of accuracy of the GeoJson file is not the best
3. GeoJson more accurate: The accuracy of the GeoJson file can be improved, for temporal and spatial complexity a small but functional version was selected.

CONCLUSIONS AND TAKEAWAYS

1. A complete, fast, and easy-to-use application was created, empowering flooding prone communities to take preventive actions, reducing the impact of floodings in their livelihood, assets, health and lifestyle.

2. The geo-referenced analysis of the flood risk data predicted by the application provides the Colombian government with a tool for early detection of emergencies, including applicability to nearby vulnerable communities.

3. The Machine Learning model developed was 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.

4. Exist a direct relation between climate variables like precipitation, temperature and atmospheric pressure that allows the assessment of flooding risk for all municipios in Colombia.

5. The best model family for this dataset was tree-based ensembles, particularly Random Forests and XGBoosts, trained with Geometric Mean of the Sensitivity as objective score, and randomly oversampled on the minority class to reduce imbalance. Empyrically, these models displayed the best performance on our training data. One possible reason could be that ensemble nature helps capture the complex relationships between the weather predictors and our target variable.

6. Future developments, like an Amber Alert system, can be created using this application as core to automatically communicate flooding risk assessments to the impacted communities via cellphone text message. Although this system might look old fashioned, is also true that text message coverage area is much wider that internet coverage, therefore the communities that might be beneficed by alerts communication through this channel is bigger.

7. The use of tools like Google BigQuery proved to be fundamental to our project, reducing the processing time by 200 times compared to SQL queries.

8. Due to the annual aggregate analysis, it was determined that flood emergencies in Colombia have occurred more frequently in the months of May and November, which is closely related to the higher levels of precipitation during the rainy season.