

PREDICTING HIGH-TENSION TRANSMISSION LINE FAILURES DUE TO ATMOSPHERIC DISCHARGES

Project Report

**TEAM 4**

*Diana Zuluaga*

*Edison Yepes*

*Camilo Gutiérrez*

*Wbeimar Ossa*

*Julián Arango*

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

Disclaimer and confidentiality

The content of this report is confidential and intended for the **Ministry of Information and Communication Technologies of Colombia**, **Correlation One Inc.**, and **Intercolombia S.A. E.S.P.** only. It is strictly forbidden to share any part of this report with any third party, without the written consent of the aforementioned organizations.

**Table of contents**

[**Introduction**](#_vgztaveq9b7z) **5**

[Business problem and business Impact](#_cw8qlyemc4lf) 5

[Our approach](#_izx6kih05t79) 5

[**Application overview**](#_9otux4d7ya84) **6**

[**Data engineering**](#_2txem5ykkzag) **8**

[Interactive front-end](#_9og5ygivnfst) 8

[Database](#_3hmo3yofw570) 8

[Solution diagram](#_z3mf2wjzghxu) 9

[**Data analysis and computation**](#_y0rk9tye3jro) **10**

[Datasets](#_xqedu8by19l5) 10

[Line Comuneros - Primavera](#_g1l6znkghd4) 10

[Line Cerromatoso - Primavera](#_giq8zevl7q1) 10

[Line La Virginia - San Carlos](#_e6lcwdlxjjrk) 11

[Data Wrangling & Cleaning](#_l4s9dewow9bg) 12

[Exploratory Data Analysis](#_gy9lstispz02) 12

[Statistical analysis & machine learning](#_9c00n4wy20c2) 18

[General regression Neural Network (GRNN): another approach we considered](#_uzxznyi0ud07) 29

[**Conclusions & future work**](#_azry7cymm08s) **31**

[**Team 4 members**](#_vvvi4gz7n55z) **32**

[**References**](#_azry7cymm08s) **33**

[**Annex 1: datasets descriptions**](#_s1rdjvuqwt4v) **35**

[Line Comuneros-Primavera](#_q4g6m2ofeoze) 35

[Line Cerromatoso-Primavera](#_1t7o4zt33va9) 36

[Line La Virginia-San Carlos](#_79zeyesadn5h) 38

# 

# Introduction

## Business problem and business Impact

ISA INTERCOLOMBIA is Colombia’s biggest player[[1]](#footnote-0) in the energy transportation market (high tension, above 220kV). The company connects energy generators with organizations distributing electrical energy to final consumers (with infrastructure capable of transporting energy at lower tensions, below 220kV). ISA INTERCOLOMBIA then focuses much of its resources into keeping their electrical transport infrastructure up and running. A transmission line failure may disrupt the delivery of electrical energy to end consumers, carrying sizable consequences, not just in direct economic terms, but also in social welfare. The company helps to provide a key resource for all people in the country. Therefore, for them it is highly valuable to anticipate those potential failures. That is the business problem we face in this project: **to improve the uptime of ISA INTERCOLOMBIA operations by timely anticipating potential failures in their tension lines**.

More specifically, the company wants to predict the failure of the Comuneros-Primavera, Primavera-Cerromatoso and San Carlos-La Virginia transmission lines due to atmospheric discharges, with at least 5 minutes of anticipation. Predictions should rely on previous atmospheric discharges information, with a true positive rate higher than 70% and a true negative rate near 80% or higher.

Making accurate predictions regarding transmission line failures will allow ISA INTERCOLOMBIA taking preventive actions to keep electrical infrastructure operational, and avoiding further electrical energy supply degradation. These predictions will enable the company to save time and financial resources associated with equipment damage prevention. Furthermore, this information would allow resuming operations faster and more efficiently than a situation in which no timely alert would be present.

Achieving this would then positively impact ISA INTERCOLOMBIA’s financial performance, as well as helping companies’ operations run smoothly, and helping to ensure national electric energy supply.

## Our approach

At the beginning, we considered the analysis of atmospheric variables from external sources relevant, even though this may result in a more complex model (e.g., training). However, during the execution of the project, we could not obtain external data to add to our solution that we were sure could help. We faced some barriers that prevented us from using open data. For example, we did not receive responses to several emails sent to IDEAM, in which we asked for guidance about the usefulness of the data they have online for open access. We also tried to find open data about climate for Colombia in other sources, but we failed.

After conducting an Exploratory Data Analysis (EDA), it was clear that the challenge the organization faces is a classification problem: given several conditions, they needed to identify whether a failure is going to happen or not within minutes; at least, the probability of a failure. We further explored which models could work. We put our attention to logistic regression, decision trees, random forests, and other models suited to this kind of problem. We also conducted a literature review to get a sense of how scientists have addressed this problem. We found that predicting these failures caused by atmospheric discharges was indeed an open problem in engineering, which many research groups are working on. As an example, we can refer to Berkowitz (2016) which tells a story about a research team testing a lighting prediction model that was built with decades of lightning historical data.

Although we reviewed more research articles, we followed closely two: Barrera (2016) and Xie, Li, Lv, & Yu, (2019). With these papers and our own conclusions from the EDA, we defined the features needed to feed several models.

In short, our approach to a suitable solutions required features of clusters of discharges (stor cells) to be created from initial data:

* Storm duration
* Discharges per minute
* Average magnitude
* Maximum magnitude
* Area
* Discharges per km2
* Distance centroid
* Distance polygon
* Maximum distance
* Failure label

# 

# Application overview

Our application can be found at [ds4a.aleatoria.co](http://ds4a.aleatoria.co). It is composed of several dashboards that allows users to explore visualizations and data related to discharges and failures over three ISA INTERCOLOMBIA transmission lines. The use cases are two: first, exploring the failure events and associated data (e.g., discharges) geographically and over time; second, predicting a failure. Users just need to visit [ds4a.aleatoria.co](http://ds4a.aleatoria.co) and go to the section they wish to use on the right-side menu.

**Exploration of previous failure events.** Users can do the following:

*General settings:*

* Select the date range to explore discharges and failures.
* Select the specific date a failure occurred.

*On a map:*

* Show a specific transmission line.
* Select the buffer (lightning corridor) over the selected transmission line.
* Show the location of discharges over the transmission line area in terms of magnitude, current or polarity.

*On a XY plot:*

* Select the time frame of discharges before a line failures over the map
* Explore the magnitude, polarity or current of discharges over time.

**Prediction of failure events given current data.** Users can do the following:

By default, in the first tab (clusters), users see indicators related to the algorithm performance (true positive rate, true negative and prediction threshold). Below, users can do the following:

* Select a transmission line.
* Select a specific failure.

That way, users can visualize discharges associated with the specific failure. Additionally, users can see the features characterizing each cluster.

To the right, users can access the second tab which points to the real-time prediction of the selected transmission line. The system triggers an alert when the prediction of any of the clusters exceeds the threshold assigned as a parameter in the algorithm. Users can visualize the features used in the prediction model, with which the line alert is generated.

# Data engineering

## Interactive front-end

The interactive front-end is hosted at DigitalOcean in a droplet (server) with basic specifications. We installed a Nginx web server to publish the application. The user interface (dashboards) was built using Python 3 and Dash (by plotly). We use a temporal domain name to make it easy for users to visit the application ([ds4a.ealetoria.co](http://ds4a.ealetoria.co)) provided by one of the team members.

The front-end consumes a database which stores all the information provided by ISA INTERCOLOMBIA (see more at [Database](#_3hmo3yofw570) section). This way, users can explore the historic data and visualizations they want to review. Furthermore, users can see real-time predictions[[2]](#footnote-1). The predictions point to five minutes after the most recent discharge event and are performed feeding the model with the most recent 24 hours of discharge data available.

As we discussed in the [Application overview](#_9otux4d7ya84) section, two use-cases were deployed: exploration and prediction. We decided to build an exploration section to provide an easy way for users to see how discharges are related to failures geographically. In some sense, this exploratory capability might help organizational efforts to improve models. And of course, the prediction section was a requirement. This section is the most important for ISA INTERCOLOMBIA as it would allow technical teams to act quickly facing an imminent failure, reducing impacts to the company and its customers.

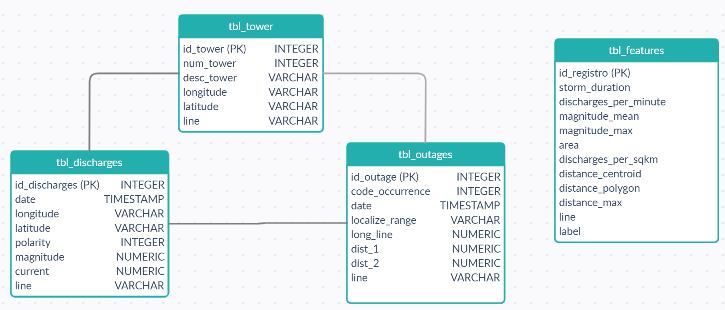
## 

## Database

We use a Postgres 12.3 database enabled on an AWS RDS instance. The instance has an access security system through security groups. We use Postgres because of its capacity, ease, available documentation and potential advantage for handling spatial data.

We built a relational model to store the base information of discharges, line failures and energy towers (*tbl\_discharges*, *tbl\_outages* and *tbl\_towers*). Additionally, we build a table of features (tbl\_features) to store the main inputs that will feed the prediction model.

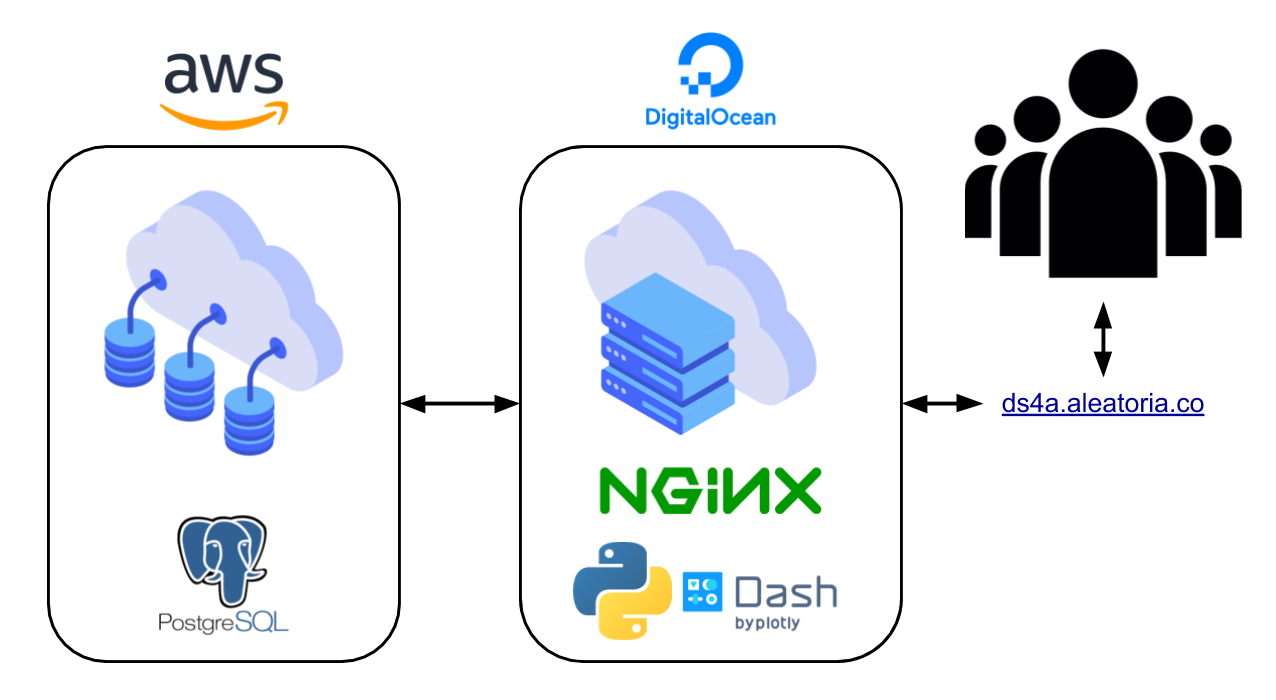
The construction and manipulation of the tables was done at all times from a Jupyter notebook, using the *ipython-sql* library.



**Figure 1. Relational diagram.**

## Solution diagram

Below, we show a high-level representation of our solution from the information technology perspective.



**Figure 2. Solution diagram.**

# 

# Data analysis and computation

## Datasets

ISA INTERCOLOMBIA provided three datasets for three transmission lines. These include historical records of atmospheric discharges near the transmission line; the historical records of transmission line outages; and the geographical location of line’s energy towers. Below we summarized the most important variables for all the data we have for each line[[3]](#footnote-2).

### Line Comuneros - Primavera

|  |  |
| --- | --- |
| Atmospheric discharge information | |
| Cases | 252.062 |
| Period (dates) | 2018-04-01 00:01:37 ---- 2019-11-30 01:56:56 (608 days) |
| Magnitude | Min: 1.8; Max: 316.7 [kA] |
| Current | Min: -316.7; Max: 156.7 [kA] |
| Transmission line failures | |
| Cases | 106 |
| Period (dates) | 2007-05-16 23:12 ---- 2020-09-16 23:24 (4873 days) |
| Location of transmission line | |
| # of energy towers | 278 |

### Line Cerromatoso - Primavera

|  |  |
| --- | --- |
| Atmospheric discharge information | |
| Cases | 10.854.274 |
| Period (dates) | 2007-01-01 00:22:25 ---- 2020-03-31 23:59:58 (4838 days) |
| Magnitude | Min: 1.2; Max: 431 [kA] |
| Current | Min: -431; Max: 354.4 [kA] |
| Transmission line failures | |
| Cases | 136 |
| Period (dates) | 2007-06-09 04:43 ---- 2020-10-20 00:14 (4883 days) |
| Location of transmission line | |
| # of energy towers | 463 |

### Line La Virginia - San Carlos

|  |  |
| --- | --- |
| Atmospheric discharge information | |
| Cases | 6.762.070 |
| Period (dates) | 2007-01-01 03:51:48 ---- 2020-03-31 23:59:56 (4839 days) |
| Magnitude | Min: 0.0; Max: 406 [kA] |
| Current | Min: -406; Max: 275 [kA] |
| Transmission line failures | |
| Cases | 156 |
| Period (dates) | 2007-01-16 16:16 ---- 2020-09-20 19:40 (4997 days) |
| Location of transmission line | |
| # of energy towers | 417 |

## Data Wrangling & Cleaning

The information provided by ISA INTERCOLOMBIA is of high quality. No major adjustments or data cleansing was necessary. A case of incomplete information corresponds to a section of the towers of the Comuneros-Primavera transmission line, which was not supplied by the company (apparently, these data points were missing). The line connects with another to form a section in which the missing energy towers were located.

To complete the missing energy towers locations, a thorough search was done visually using Google Maps, identifying the Primavera power substation located in the municipality of Puerto Berrío. From the power substation we followed the transmission line by the shadow over the map until the energy tower for which we received location information.



**Figure 3. Example of energy tower search.**

Another adjustment we made was calculating the location of the faults over the transmission line. The original dataset indicated the distance of each fault from each of the substations that make up the line. Taking these distances, the geographic location of the faults was calculated. For missing cases, an imputation was made.

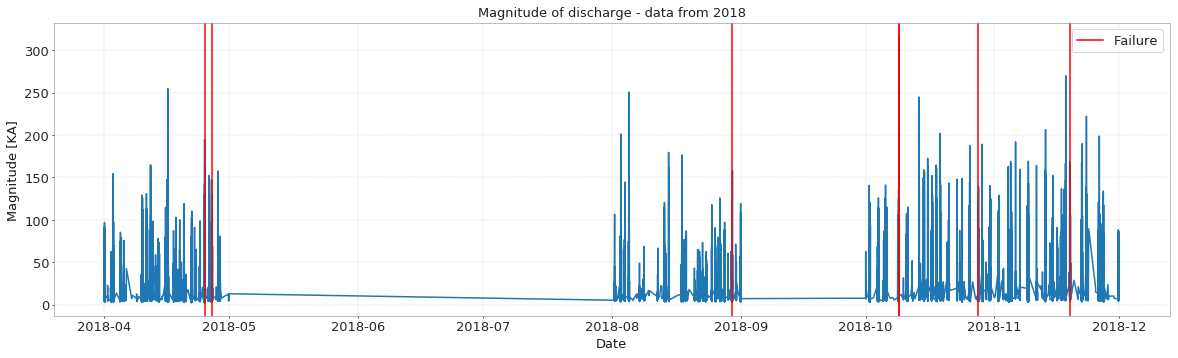
## Exploratory Data Analysis

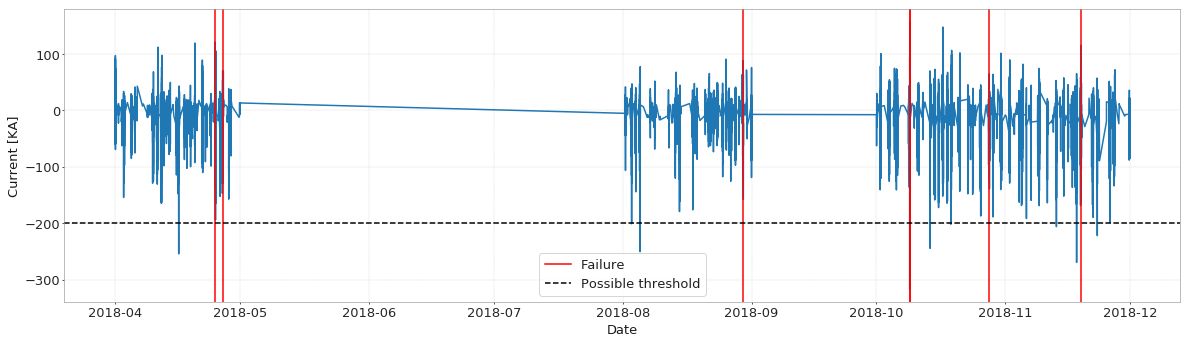
**We explored how lightning was associated in time and space with failures. Some hypotheses are outlined below:**

* **Groups of discharges near the transmission lines could be associated with line failures.**
* **The magnitude of discharges is relevant for line failures.**
* **Other climate information could be used to predict groups of discharges. At the moment, a lightning cannot be predicted individually[[4]](#footnote-3).**

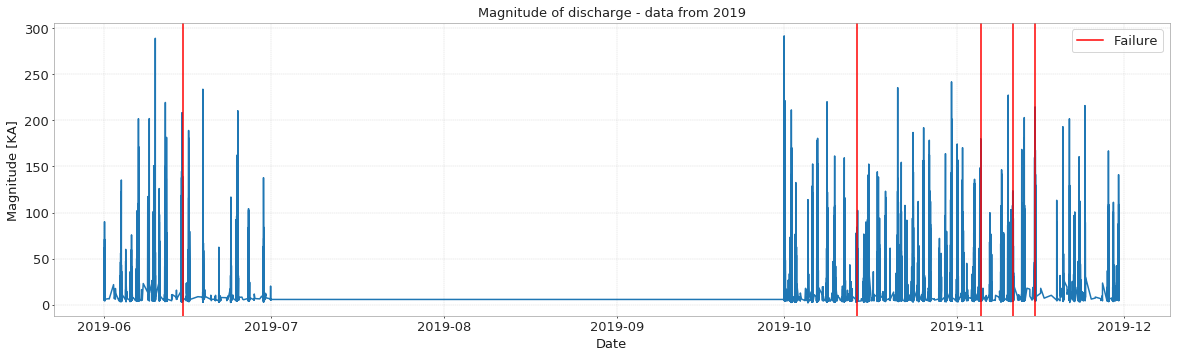
**Some plots below reveal some association between lighting and specific failures. Although it is not very clear, we see that failures seem to appear with more frequent previous discharges.**

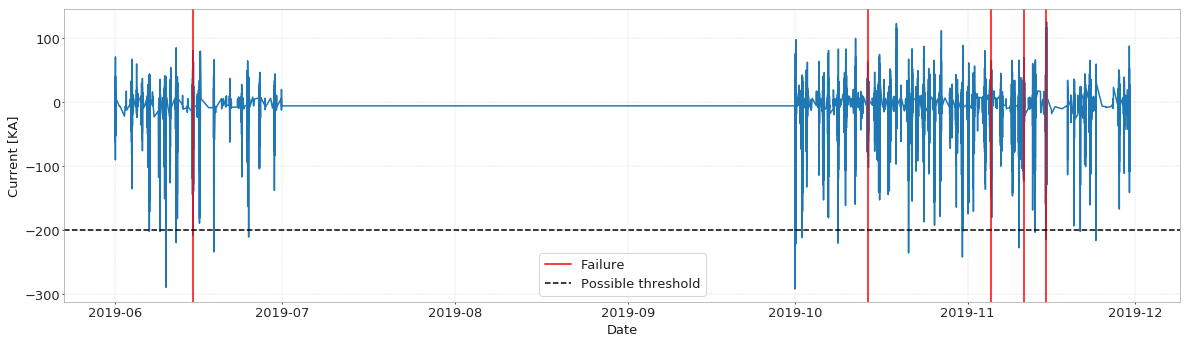
**Lightnings (magnitude and current) and failures**





**Figure 4. Comuneros-Primavera line. Available data 2018.**



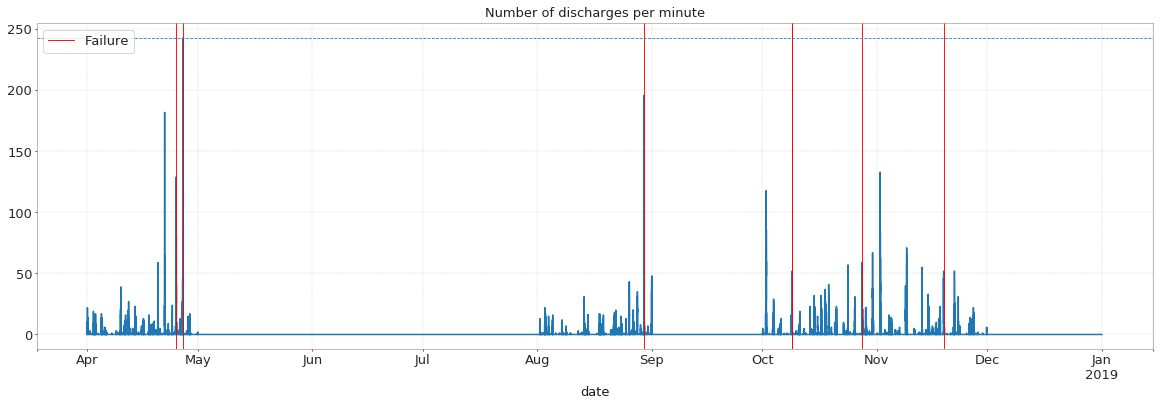


**Figure 5. Comuneros-Primavera line. Available data 2019.**

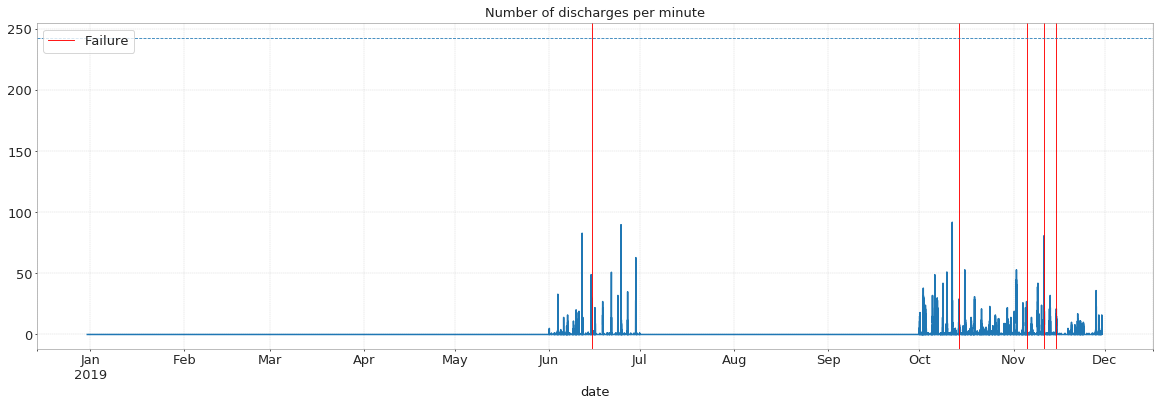
**A detailed exploration of the discharge information was developed, crossing with the information of the lines' operating outputs on a temporary basis, looking for a pattern in the discharges and their characteristics that could take the line out of operation. One of the main approaches explored was to analyze the magnitude and current of the discharges per month, day and minute.**

**Lightnings per minute and failures**

**Additionally, we calculated a density metric to assess whether more discharges in shorter periods of time are with failures.**



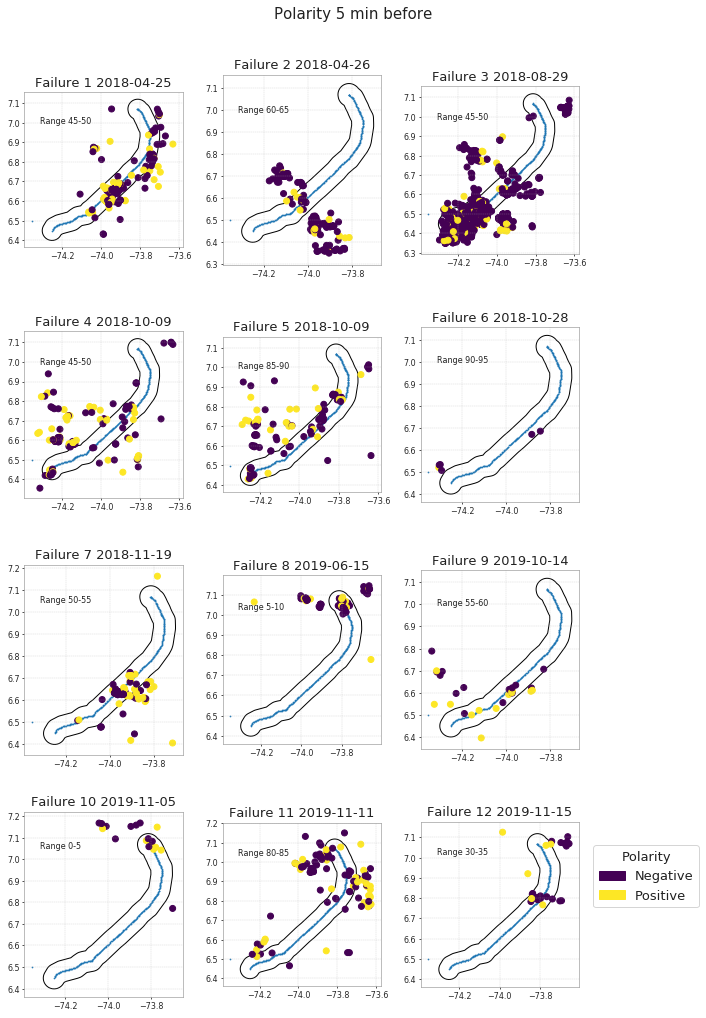
**Figure 6. Comuneros-Primavera line. Available data 2018.**



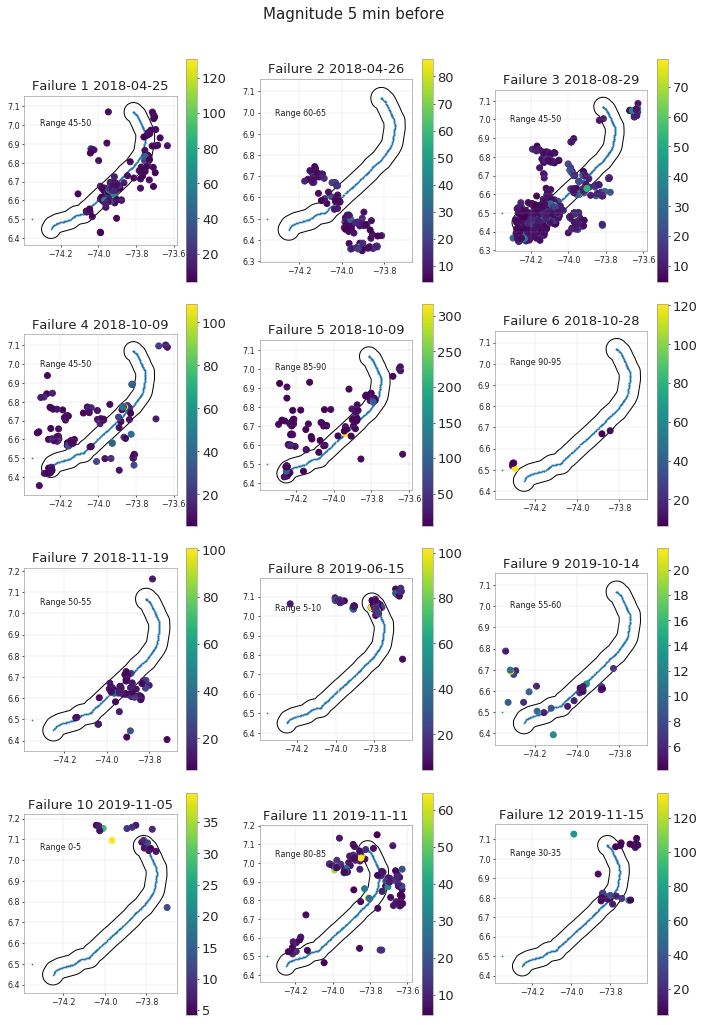
**Figure 7. Comuneros-Primavera line. Available data 2019.**

**Based on this analysis, we hypothesize a direct relationship, not only spatial but also temporal, between these groups of lightning bolts and the faults; these storm cells can allow us to model the phenomenon with a suitable window of time and space.**

**In the maps below, we can see groups of discharges close to the moment of failure. This suggests that grouping or clustering of discharge could be a good proxy to predict failures.**



**Figure 8. Spatial location of discharges near to the transmission line: magnitude**



**Figure 9. Spatial location of discharges near to the transmission line: polarity**

**Once the information had been analyzed, it became clear that it was a spatio-temporal problem where groups of downloads can be identified at a given time and space. From this analysis, the features are built to feed the clustering model.**

## Statistical analysis & machine learning

After we finished the exploratory data analysis, we started investigating possible solutions for the problem. One approach seemed particularly well suited: the one carried out by Luisa Fernanda Barrera in her master's thesis called “*Probability* *Estimation of operational failure in transmission lines due to atmospheric electric shocks*" (Barrera, 2016).

In the study, a clustering of atmospheric discharges is proposed to identify “storm cells”. These storm cells show similar characteristics shortly before the transmission line failure caused by a particular discharge. Although we found it coherent to apply this clustering methodology, no predictive algorithms were used in the study to calculate the probability of failure. A deterministic approach was employed, a similarity metric, matching the features of a given storm cell to the features of previous storm cells linked to a failure.

Taking this clustering idea for our project, we used the transmission lines data given by ISA INTERCOLOMBIA **(Comuneros-Primavera, Primavera-Cerromatoso, San Carlos-La Virginia)** and we applied the ST-DBSCAN clustering algorithm (Birant & Kut, 2007). We consider it as the most suitable algorithm to perform the clustering, as the problem presents spatial and time dimensions. Opposed to existing algorithms that perform density-based clustering, this algorithm is able to discover spatial clusters through the time dimension of such objects, relying on the DBSCAN algorithm capabilities, which has the ability to discover clusters with arbitrary shapes.

It is important to note that this algorithm has the ability to process large volumes of data and does not require the number of clusters to be created in advance.

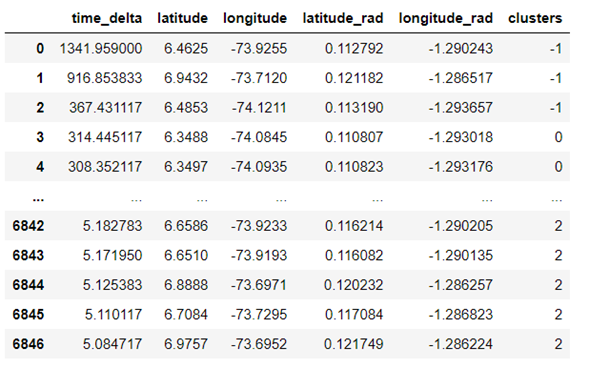
The algorithm is powered by spatial and temporal information, and three parameters, known as EPS1, EPS2 and Min\_samples should be provided. EPS1 is the value used to measure geographical proximity (latitude and longitude) and EPS2 is the parameter used to measure the similarity of temporal values. With these parameters, space-time information is selected to build point neighborhoods that meet both criteria: groups of points that are spatially close and observed in consecutive time spans according to the selected time unit.

The algorithm also requires defining the Min\_samples parameter which is defined as the minimum of points as input to form a cluster, to avoid generating small-sized clusters since these could be very dense in time-space and therefore contain little information.

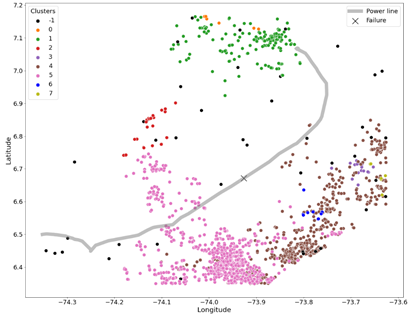
One of the main advantages of this algorithm is that in case a change is needed in the space and time criteria to create clusters, it is only required to modify the distance and time parameters. Afterward, testing different combinations should be performed to improve the predictive capacity of a possible model that is based on clustering information thrown.

We decided to perform the clustering by using 24 hours of discharges data prior to a failure. This design decision was made to remove information from days free of failures. Not doing so would generate a large number of clusters that cannot be directly associated with a failure event. Furthermore, this allows to improve the issue of data imbalance from the original datasets. Additionally, discharges data from five minutes prior to a failure is omitted too in order to obtain the attributes of the storm cells with an adequate time window, as requested by ISA INTERCOLOMBIA.

Given our conclusions from the EDA, the information obtained from ISA INTERCOLOMBIA during project meetings and and the literature review we conducted, we decided to perform the clustering with 10 kilometers away, a time window of 10 minutes, as well as a parameter of 5 minimum points for clustering. The results obtained are presented in the figures and tables below.

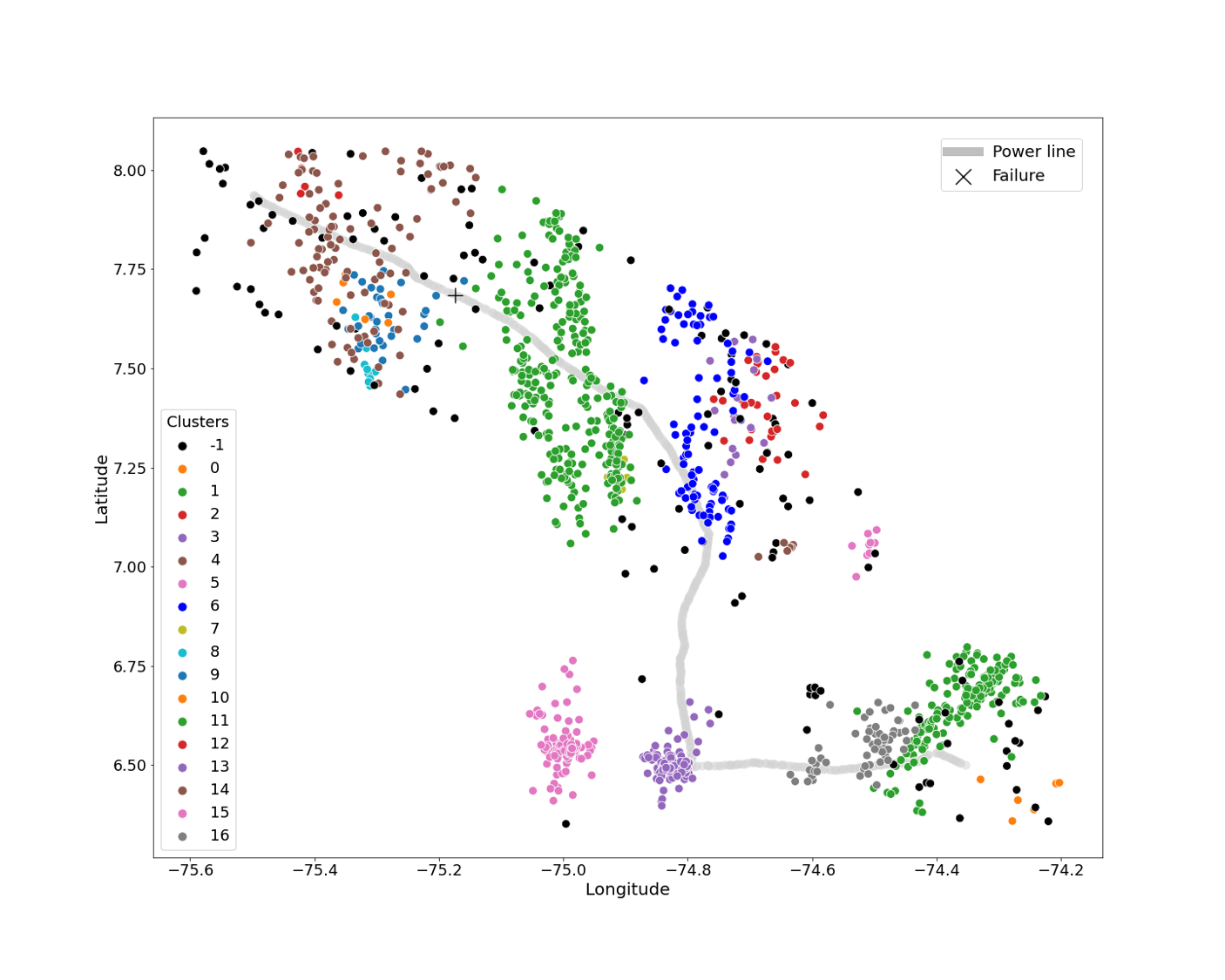


**Table 1. Example of clustering of the Comuneros-Primavera line.**

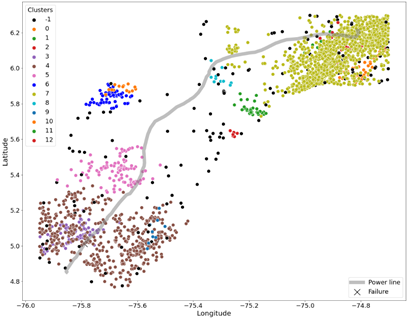


**Figure 10. Clustering of the Comuneros- Primavera line.**

The figures show the clusters created, labeled with a number and a color scale. Those that are labeled with -1 and black color are discharges that cannot be grouped in space-time; that is, they are considered noise within our clustering model.



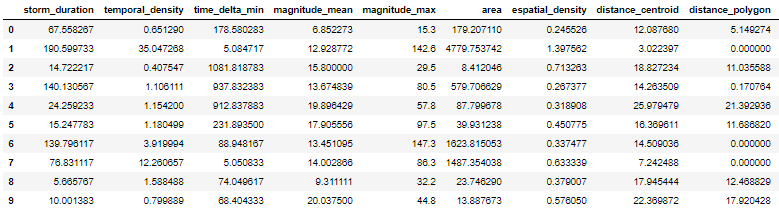
**Figure 11. Clustering of the Primavera – Cerromatoso line.**



**Figure 12. Clustering of the San Carlos – La Virginia line.**

After the clustering process for the three transmission lines, the next step is to build the features of the resulting clusters. This way, we can associate storm cells features to failures as well as associate storm cells features to normal operation, looking for the storm cell that has the closest discharge in space and time to the fault. This is precisely the data we need to train the models under consideration and test their predictive ability.

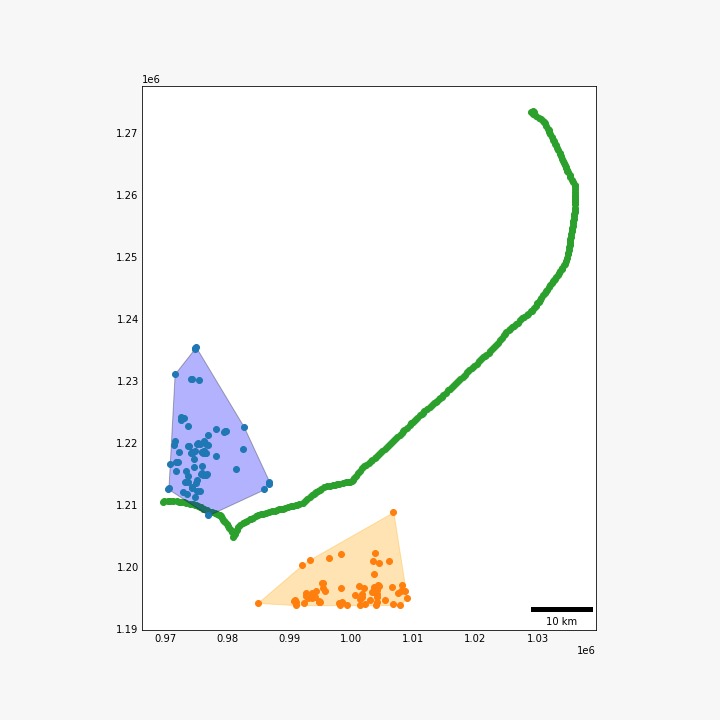
To this end, the key features of storm cells (clusters) were selected from the studies by Barrera (2016), Xie et al. (2019), Sun, Wang, & Zheng (2019), and Morales, Orduña & Rehtanz (2014), as well as features proposed by team members, which were considered relevant and feasible to build in accordance to the information provided by ISA INTERCOLOMBIA. The description of these features can be found below:



**Table 2. Example of calculated features of Comuneros- Primavera line.**

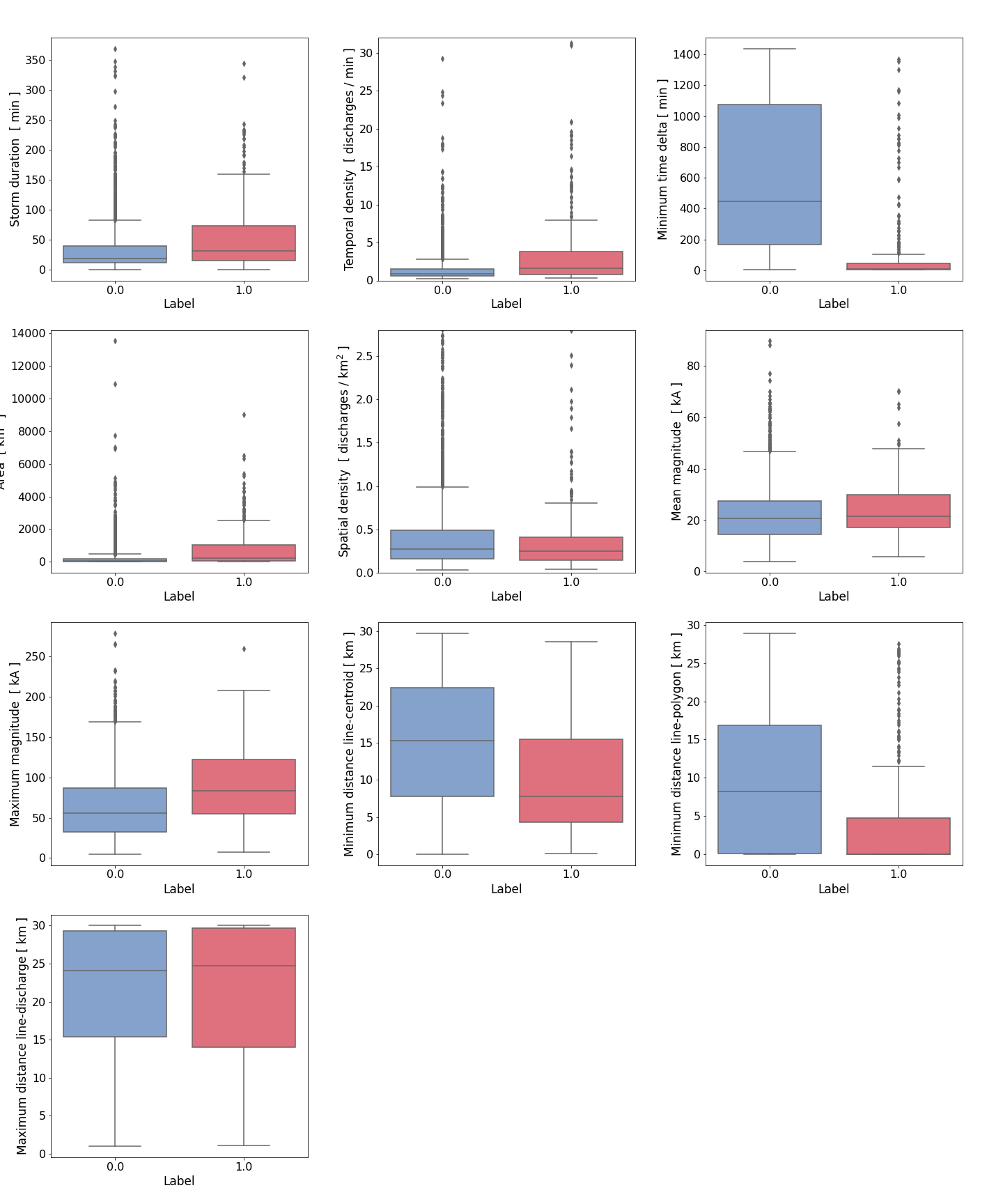
|  |  |
| --- | --- |
| **Feature** | **Description** |
| Storm duration | Duration of storm cell. [Minutes]. |
| Temporal density | Quotient between the number of discharges that occurred in the time window and the time spanned by the storm cell (cluster). [lightning strikes / minute]. |
| Minimum time delta | Time in minutes between the last cluster download and the failure. The last discharge occurred in a cluster. |
| Mean magnitude | Mean magnitude of the lightning in the storm cell (cluster). [kA]. |
| Maximum magnitude | Maximum magnitude of the lightning in the storm cell (cluster). [kA]. |
| Area | Area covered by the storm cell (cluster). This measurement covers the discharges of the selected time window. [km2] |
| Espatial density | Density of discharges within the storm cell (cluster). Quotient between the number of discharges and the area of the storm cell. [lightning strikes / km2] |
| Distance centroid | Minimum distance from the transmission line to the geometric center of the storm cell (cluster). [Meters]. |
| Distance polygon | Minimum distance from the transmission line to XXXXXXX. [Meters]. |
| Maximum distance | Minimum distance from the transmission line to the furthest point from the storm cell (cluster). [Meters]. |
| Failure Label | Binary label indicating which storm cell is considered the cause of the transmission line failure. Assigned by locating in which storm cell (cluster) the closest discharge in time and space to the failure is. |

**Table 3. Theoretical Key Features.**



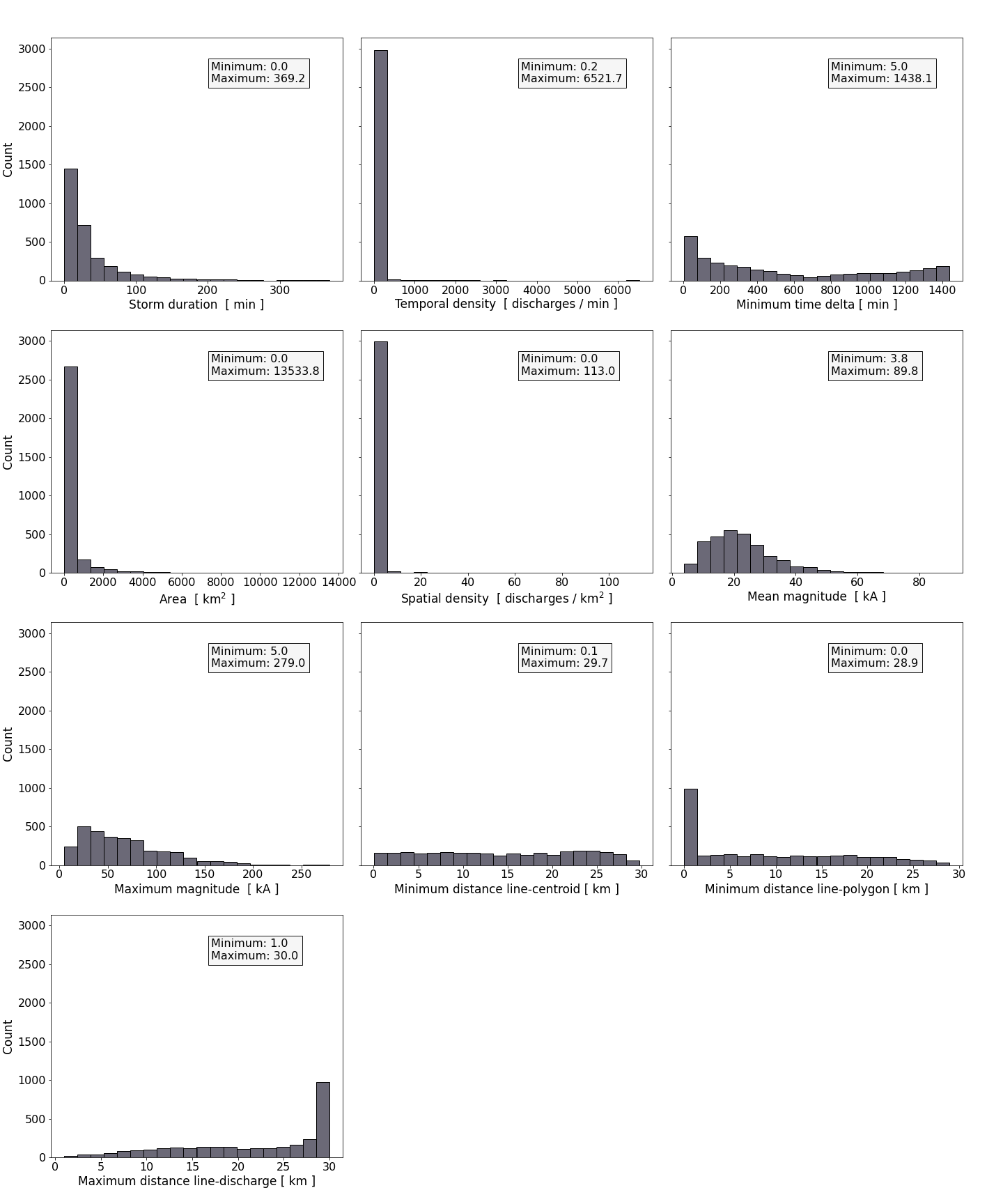
**Figure 13. Example of the construction of Area key Features.**

We compared the feature distributions by label (failure, no failure) across clusters (see the next figure). For most of the features it can be seen a difference between “failure” storm cells and “normal operation” storm cells.



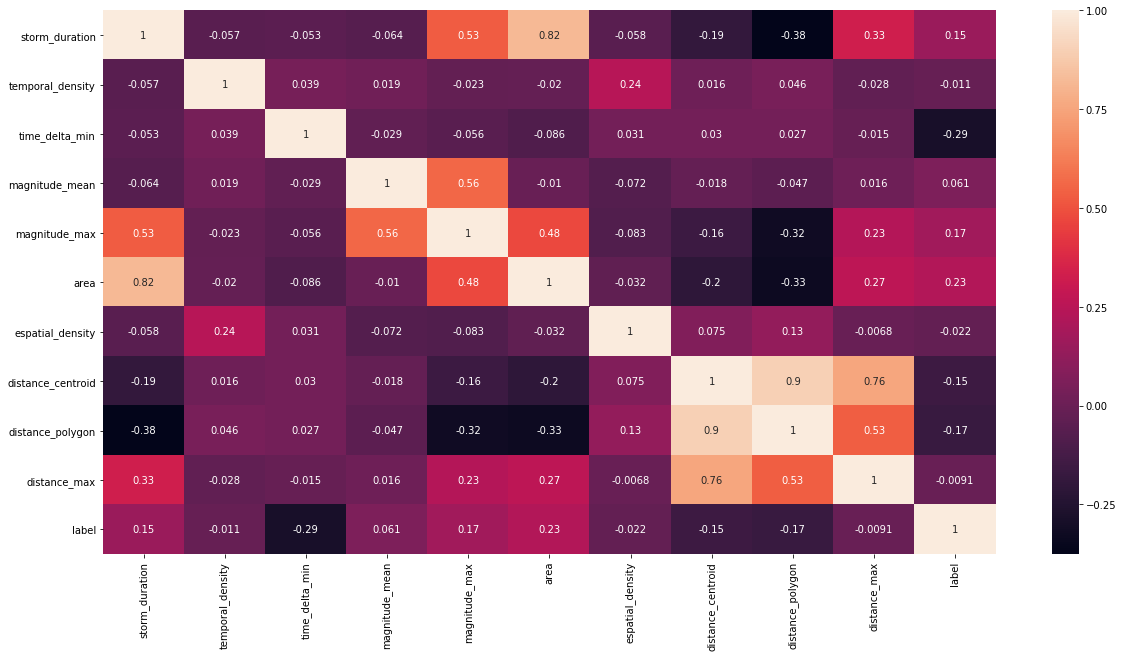
**Figure 14. Boxplot of key features.**

It can also be observed that key features mostly have a bias to the left, because there is a greater number of storm cells with a short duration and a small area, a characteristic that is present in variables derived from these two metrics. Other key features related to magnitudes and distances have a more uniform behavior; this can be seen the following figure:



**Figure 15. histogram of key features.**

We ran an analysis to check which variables were highly correlated and identify possible effects due to multicollinearity. After testing removing one or the other, the minimum distance of the cluster was taken out from the analysis as no predicted power was lost.

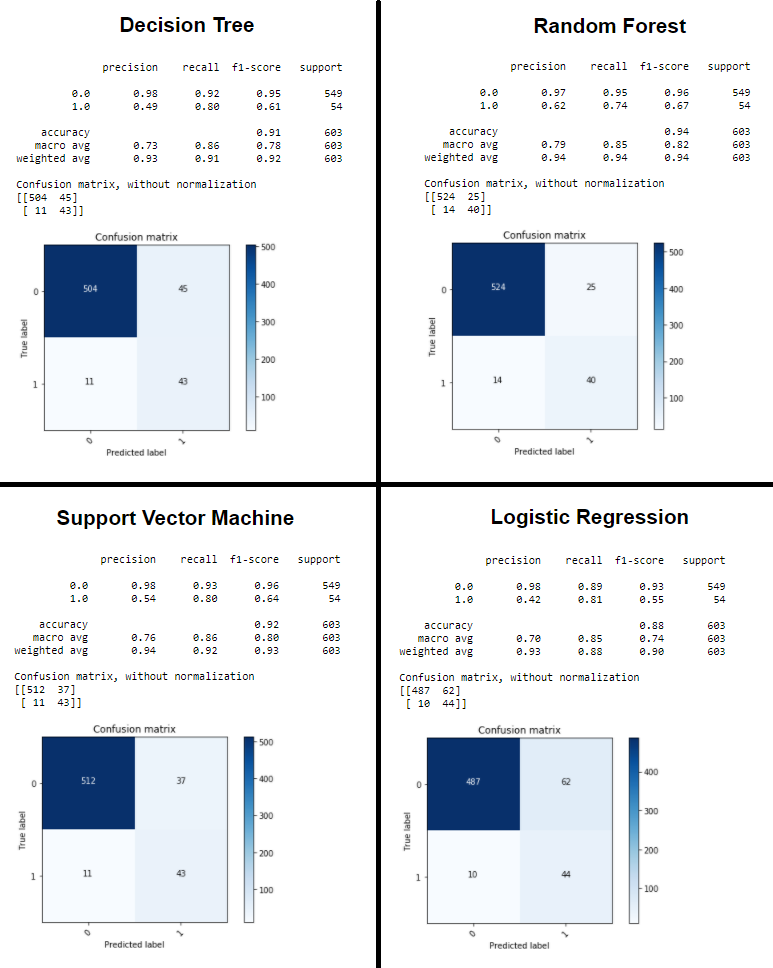


**Figure 16. Correlation heatmap of key features.**

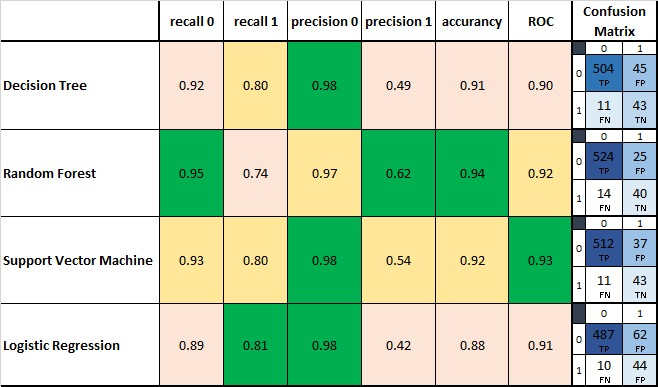
We analyzed which approach was better suited to predict the transmission line failures. We concluded that machine learning techniques are adequate options to solve the challenge, as it is a binary classification problem which ought to identify which storm cells are associated with line failures. Furthermore, based on this classification, obtain a prediction of which future storm cell would cause a failure.

For this purpose, we used the features we discussed previously as input into different machine learning models like logistic regression, support vector machines, decision trees and k-nearest neighbors. We wanted to compare each model’s prediction performance and check whether one or more of these models meet the expectations of ISA INTERCOLOMBIA: true positive (TP) rate of 70% or higher and a true negative (TN) rate of 80%. The latter is very important, as a false prediction of failure generates additional costs due to the resources devoted to timely addressing the failure.

The above models were tested with parameters set to use balance weights to achieve data balancing, in order to take into account that the proportion between the storm cells associated to failures and storm cells associated to normal operation is unbalanced, since more storm cells are associated to normal operation. The results of the confusion matrix, the ROC curve and the metrics requested by ISA INTERCOLOMBIA were taken into account to choose the most appropriate estimation method. The results of these three metrics can be seen below:

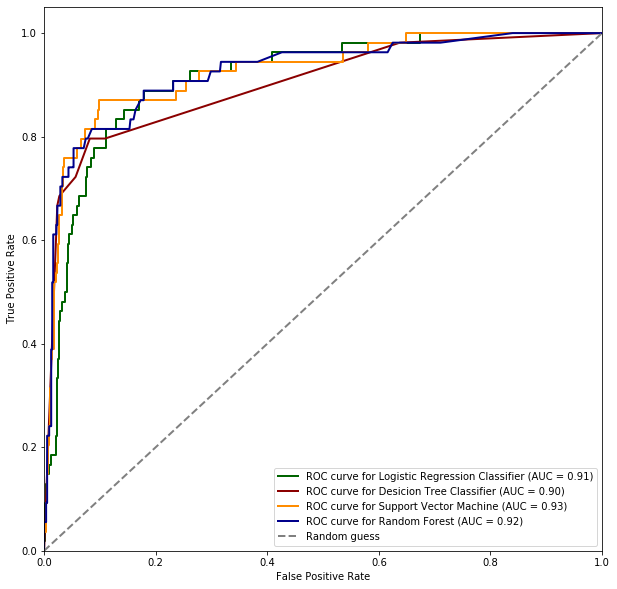


**Figure 17. Confusion matrices and precision metrics for the tested models.**

****

**Table 4. Confusion matrices summary (TP,TN,FP,FN)**.

As can be seen, the Support Vector Machine model is the one that shows a better proportion in the rate of true negatives and true positives, achieving 80% and 93% respectively. It can also be seen in the ROC curve that the support vector machine model presents an AUC indicator of 93% which indicates that it is a better model. Although the Random Forest presents similar rates and a ROC curve with an AUC indicator of 92%, this was calculated making the probability threshold with which the storm cells are classified much more acidic. Because of this, the support vector machine model is much more parsimonious and is the one we consider most suitable to make the prediction of a one line failure due to a storm cell for any of the three lines in this case.



**Figure 18. ROC Curve for the tested models.**

### General regression Neural Network (GRNN): another approach we considered

We explored the proposal made by Xie et al. (2019). The paper uses a generalized regression neural network (GRNN) to predict transmission line outages given the limitations from the data the researchers gathered. We followed all the steps outlined in the study by coding in R (R version 4.0.2 (2020-06-22) -- "Taking Off Again") using the package yager[[5]](#footnote-4). In short, these are the steps we followed:

1. **Clustering**. The lightning must be divided into groups according to time. Using the time of first lightning discharge, the lightning discharges are grouped every 15 minutes.
2. **Construct the features set**. These are calculated by the lightning data in every group and the location of the transmission tower. The output data represent the occurrence of lightning failure in the next time period (if an outage occurs, yi =1,or yi =0). The features are:
   1. the number of discharges in the lightning corridor,
   2. the nearest distance between the line and the discharge,
   3. the average discharge magnitude in the lightning corridor,
   4. and the peak current of the nearest lightning strike.
3. **Choose the train and test samples from the whole features set**. Because the number of line failure cases is much less than the number of normal operation cases, the “no failure” cases are randomly selected from the normal operation features to form a balanced training and test set. All the failure cases are all selected. 80% of the total sample is selected as the training set; the remaining 20% is the test set.
4. **Optimize the smooth parameter sigma** (the only hyperparameter).
5. **Use the trained GRNN to predict line failures from discharges in the test samples**. After obtaining the prediction sample output y0, the study takes 0.5 as a threshold to classify the sample. If y0 is greater than 0.5, the sample is classified as a line failure case.
6. **Check the effectiveness of predicted cases**. Compare the prediction status of the transmission line with the recorded data to confirm the accuracy of prediction.

We were able to train the model, but we ran into trouble in two aspects: finding the optimal smooth parameter and obtaining proper predictions. The second issue was more pervasive as the model generated NaN values (NaN means ‘Not a Number’).

Unfortunately, we did not have the time to further explore this model.

# Conclusions & future work

In this project, we aimed to predict transmission line failures due to lightning strikes. We could confirm that groups or clusters of discharges (strom cells), as well as some of its features (mean magnitude, peak magnitude, number, distance, etc.) are indeed associated with line failures. Computational thinking mixed with the use of machine learning algorithms are adequate tools to address this classification problem. Moreover, the approach used tackles a key component of the problem which is the time dimension, by means of the clustering strategy (ST-DBSCAN). We were able to find to what extent some models performed to predict line failures and the objectives pursued by ISA INTERCOLOMBIA have been met. We believe there is more that can be done; our final product can be seen as a baseline for future developments given its proper theoretical support which would allow further contributions.

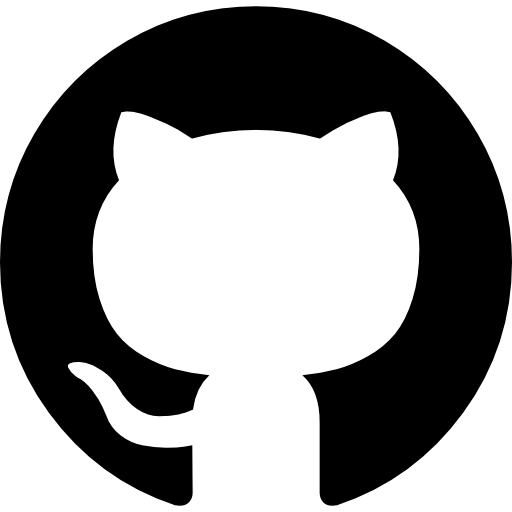
We got to know the challenges of building a good solution for a problem like the one we were assigned. It is just not about a dataset and statistical software. It is about information technology (IT) infrastructure ready to serve any user at the right moment. We devoted a great deal of effort to learn how to use IT to ensure the solution could be accessed, that it could deliver value for real.

Additionally, we learned that researchers around the world are developing more complex models, leveraging different meteorological data to improve predictions of lightning areas. According to the European Centre for Medium-Range Weather Forecasts (ECMWF) “*The main reason why we can predict lightning is that it is linked to particular weather conditions*[[6]](#footnote-5)*.*” This tells us that indeed, considering other variables would allow us to generate better predictions.

Future work is promising. As Barrera (2016) and Xie et al. (2019) suggested in their particular research, two clear paths should be explored:

1. Include more meteorological or climate data into consideration. We know this is feasible, but open data repositories seemed not to exist for our purposes. Future researchers should focus on getting more data related to the lightning discharges, generate some hypotheses and test them.
2. Other features of storm cells could be calculated. For instance, what about the cloudiness data? What about temperature? Even without additional variables, there might be other features that could be associated with the failures.

# Team 4 members

[github](https://github.com/team4-DS4A)

|  |  |
| --- | --- |
|  | **Diana Patricia Zuluaga Pulgarín**  Environmental engineer; MSc Water Resources Engineering student  [**Social icon**](https://www.linkedin.com/in/diana-zuluaga-pulgarin/) |
|  | **Edison Javier Yepes Sánchez**  Business Intelligence Consultant  [**Social icon**](https://www.linkedin.com/in/edison-yepes) |
|  | **Camilo Gutiérrez Ramírez**  Civil engineering student  [**Social icon**](https://www.linkedin.com/in/cagutierrezra/) |
|  | **Wbeimar Ossa Giraldo**  Economist; Finance specialist; MA Financial management student  [**Social icon**](https://www.linkedin.com/in/wbeimarossa) |
|  | **Julián Arango Ochoa**  Behavioral lead  [**Social icon**](https://www.linkedin.com/in/jarangoo/) |

# References

Barrera, L. F. (2016) **Estimación de probabilidad de falla operacional en líneas de transmisión debido a descargas eléctricas atmosféricas**. [Master 's thesis, Universidad Nacional de Colombia - Sede Medellín]. Institutional Repository Universidad Nacional de Colombia. Retrieved from <http://www.bdigital.unal.edu.co/55077/1/1116723229.2016.pdf>

Berkowitz, R. (2016) Can We Predict Lightning? **Scientific American**. Retrieved from <https://www.scientificamerican.com/article/can-we-predict-lightning/>

Birant, D., & Kut, A. (2007). **ST-DBSCAN: An algorithm for clustering spatial–temporal data**. **Data & knowledge engineering**, *60*(1), 208-221.

Dokic, T., Pavlovski, M., Gligorijevic, D., Kezunovic, M., & Obradovic, Z. (2019, January). **Spatially aware ensemble-based learning to predict weather-related outages in transmission**. In **Proceedings of the 52nd Hawaii International Conference on System Sciences**.

Morales, J. A., Orduña, E., & Rehtanz, C. (2014). Classification of lightning stroke on transmission line using multi-resolution analysis and machine learning. **International Journal of Electrical Power & Energy Systems**, 58, 19-31.

Petrović, I., Nikolovski, S., Baghaee, H. R., & Glavaš, H. (2019). Determining Impact of Lightning Strike Location on Failures in Transmission Network Elements Using Fuzzy Decision-Making. **IEEE Systems Journal**, 14(2), 2665-2675.

Sun, C., Wang, X., & Zheng, Y. (2019). Data-driven approach for spatiotemporal distribution prediction of fault events in power transmission systems. **International Journal of Electrical Power & Energy Systems**, 113, 726-738.

Xie, Y., Li, C., Lv, Y., & Yu, C. (2019). **Predicting lightning outages of transmission lines using generalized regression neural network**. **Applied Soft Computing Journal**, *78*, 438-446.

# Annex 1: datasets descriptions

Below we describe all datasets in tables[[7]](#footnote-6).

## Line Comuneros-Primavera

**Table 1.** Atmospheric discharge information. *Cases/observations*: 252.062

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Scale /Range** | **Description** |
| date | Datetime:  YYY-MM-DD HH:MM:SS | Min: 2018-04-01 00:01:37  Max: 2019-11-30 01:56:56  Range: 608 days | Detailed Date and hour of the lightning |
| longitude | Type: float  Units: degrees  X decimal places. | Max: -73.6245  Min: -74.3452 | Estimated location of lightning  (longitude) |
| latitude | Type: float  Units: degrees  X decimal places. | Max: 7.1682  Min: 6.3485 | Estimated location of lightning  (latitude) |
| polarity | Type: integer  Categorical | -1, 1 | The electric potential at the ends of a circuit that indicates the direction of electric flow, it can be positive or negative. |
| magnitude | Type: float  Units: kiloampere | Min: 1.8  Max: 316.7 | The electromagnetic force of the lightning between electrical conductors |
| current | Type: float  Units: kiloampere | Min: -316.7  Max: 156.7 | Rate of flow of [electric charge](https://en.wikipedia.org/wiki/Electric_charge) past a point. Equals Polarity per magnitude. |

**Table 2.** Comuneros-Primavera transmission line failures. *Cases/observations*: 106

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Scale/Range** | **Description** |
| date | Datetime:  YYY-MM-DD HH:MM | Max: 2020-09-16 23:24  Min: 2007-05-16 23:12  Range: 4873 Days | Detailed date and hour of the failure |
| dist\_comuneros | Type: float  Units: km | Max: 112.3  Min: 0.0 | Distance from energy tower to Comuneros power substation |
| dist\_primavera | Type: float  Units: km | Max: 110.5  Min: 0.0 | Distance from energy tower to Primavera power substation |
| R\_inf | Integer | Max: 90  Min: 0  Range: 90 | Lower boundary, in percentage across the line, where the failure happened. |
| R\_Sup | Integer | Max: 95  Min: 5  Range: 90 | Upper boundary, in percentage across the line, where the failure happened. |

**Table 3.** Location of Comuneros-Primavera transmission line. *Cases/observations*: 242

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Scale/Range** | **Description** |
| tower | Type: integer | Max: 278  Min: -0 | Number of energy towers |
| description | Type: string | NA | Description of energy towers |
| longitude | Type: float  Unit: degrees  14 decimal places. | Max: -74.35224  Min: -73.75105  Range: 0.60118 | Location of energy towers (longitude) |
| latitude | Type: float  Unit: degrees  14 decimal places. | Max: 6.4484025  Min: 7.0689245  Range: 0.62052 | Location of energy towers (latitude) |

## 

## Line Cerromatoso-Primavera

**Table 4.** Atmospheric discharge information. *Cases/observations*: 10.854.274

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Scale /Range** | **Description** |
| date | Datetime:  YYY-MM-DD HH:MM:SS | Min: 2007-01-01 00:22:25  Max: 2020-03-31 23:59:58  Range: 4838 days | Detailed Date and hour of the lightning |
| longitude | Type: float  Units: degrees  X decimal places. | Max: -74.2  Min: -75.6  Range: 0.7198 | Estimated location of lightning  (longitude) |
| latitude | Type: float  Units: degrees  X decimal places. | Max: 8.050  Min: 6.35  Range: 0.8197 | Estimated location of lightning  (latitude) |
| polarity | Type: integer  Categorical | -1, 1 | The electric potential at the ends of a circuit that indicates the direction of electric flow, it can be positive or negative. |
| magnitude | Type: float  Units: kiloampere | Min: 1.2  Max: 431 | The electromagnetic force of the lightning between electrical conductors |
| current | Type: float  Units: kiloampere | Min: -431  Max: 354.4 | Rate of flow of [electric charge](https://en.wikipedia.org/wiki/Electric_charge) past a point. Equals Polarity per magnitude. |

**Table 5.** Outages in the Cerromatoso-Primavera transmission line. *Cases/observations*: 136

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Scale/Range** | **Description** |
| outage\_date | Datetime:  YYY-MM-DD HH:MM | Max: 2020-10-20 00:14  Min: 2007-06-09 04:43  Range: 4883 Days | Detailed date and hour of the failure |
| dist\_cerromatoso | Type: float  Units: km | Max: 238.9  Min: 0.0 | Distance from energy tower to Cerromatoso power substation |
| dist\_primavera | Type: float  Units: km | Max: 243.6  Min: 0.0 | Distance from energy tower to Primavera power substation |
| R\_inf | Integer | Max: 90  Min: 0  Range: 90 | Lower boundary, in percentage across the line, where the failure happened. |
| R\_Sup | Integer | Max: 95  Min: 5  Range: 90 | Upper boundary, in percentage across the line, where the failure happened. |

**Table 6.** Location of Cerromatoso-Primavera transmission line. *Cases/observations*: 463

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Scale/Range** | **Description** |
| tower | Type: integer | Max: 404  Min: 0 | Number of energy towers |
| description | Type: string | NA | Description of energy towers |
| longitude | Type: float  Unit: degrees  14 decimal places. | Max: -74.35  Min: -75.5  Range: 0.60118 | Location of energy towers (longitude) |
| latitude | Type: float  Unit: degrees  14 decimal places. | Max: 7.938  Min: 6.487  Range: 0.62052 | Location of energy towers (latitude) |

## Line La Virginia-San Carlos

**Table 7.** Atmospheric discharge information. *Cases/observations*: 6.762.070.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Scale /Range** | **Description** |
| date | Datetime:  YYY-MM-DD HH:MM:SS | Min: 2007-01-01 03:51:48  Max: 2020-03-31 23:59:56  Range: 4839 days | Detailed Date and hour of the lightning |
| longitude | Type: float  Units: degrees  X decimal places. | Max: -71.07  Min: -78.43  Range: 0.7198 | Estimated location of lightning  (longitude) |
| latitude | Type: float  Units: degrees  X decimal places. | Max: 12.443  Min: 0.1447  Range: 0.8197 | Estimated location of lightning  (latitude) |
| polarity | Type: integer  Categorical | -1, 1 | The electric potential at the ends of a circuit that indicates the direction of electric flow, it can be positive or negative. |
| magnitude | Type: float  Units: kiloampere | Min: 0.0  Max: 406 | The electromagnetic force of the lightning between electrical conductors |
| current | Type: float  Units: kiloampere | Min: -406  Max: 275 | Rate of flow of [electric charge](https://en.wikipedia.org/wiki/Electric_charge) past a point. Equals Polarity per magnitude. |

**Table 8.** Outages in the La Virginia-San Carlos transmission line. *Cases/observations*: 156

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Scale/Range** | **Description** |
| outage\_date | Datetime:  YYY-MM-DD HH:MM | Max: 2020-09-20 19:40  Min: 2007-01-16 16:16  Range: 4997 Days | Detailed date and hour of the failure |
| dist\_lavirginia | Type: float  Units: km | Max: 220.7  Min: 0.0 | Distance from energy tower to La Virginia power substation |
| dist\_sancarlos | Type: float  Units: km | Max: 207.32  Min: 0.0 | Distance from energy tower to San Carlos power substation |
| R\_inf | Integer | Max: 90  Min: 0  Range: 90 | Lower boundary, in percentage across the line, where the failure happened. |
| R\_Sup | Integer | Max: 95  Min: 5  Range: 90 | Upper boundary, in percentage across the line, where the failure happened. |

**Table 9.** Location of La Virginia-San Carlos transmission line. *Cases/observations*: 417

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Scale/Range** | **Description** |
| tower | Type: integer | Max: 416  Min: 0 | Number of energy towers |
| description | Type: string | NA | Description of energy towers |
| longitude | Type: float  Unit: degrees  14 decimal places. | Max: -74.81  Min: -75.85  Range: 0.60118 | Location of energy towers (longitude) |
| latitude | Type: float  Unit: degrees  14 decimal places. | Max: 6.213  Min: 4.853  Range: 0.62052 | Location of energy towers (latitude) |

**Figure List**

Figure 1. Relational diagram. pag 9

Figure 2. Solution diagram. pag 9

Figure 3. Example of energy tower search. pag 12

Figure 4. Comuneros-Primavera line. Available data 2018. pag 13

Figure 5. Comuneros-Primavera line. Available data 2019. pag 14

Figure 6. Comuneros-Primavera line. Available data 2018. pag 14

Figure 7. Comuneros-Primavera line. Available data 2019. pag 15

Figure 8. Spatial location of discharges near to the transmission line: magnitude. pag 16

Figure 9. Spatial location of discharges near to the transmission line: polarity. pag 17

Figure 10. Clustering of the Comuneros- Primavera line. pag 20

Figure 11. Clustering of the Primavera – Cerromatoso line. pag 20

Figure 12. Clustering of the San Carlos – La Virginia line. pag 21

Figure 13. Example of the construction of Area key Features. pag 23

Figure 14. Boxplot of key features. pag 24

Figure 15. histogram of key features. pag 25

Figure 16. Correlation heatmap of key features. pag 26

Figure 17. Confusion matrices and precision metrics for the tested models. pag 27

Figure 18. ROC Curve for the tested models. pag 29

**Tables List**

Table 1. Example of clustering of the Comuneros-Primavera line. pag 19

Table 2. Example of calculated features of Comuneros- Primavera line. pag 21

Table 3. Theoretical Key Features. pag 22

Table 4. Confusion matrices summary (TP,TN,FP,FN). pag 28

1. 71% of market size according to ISA INTERCOLOMBIA <http://www.isaintercolombia.com/Paginas/67/transmision-de-energia-electrica> [↑](#footnote-ref-0)
2. Naturally, lightning data should be loaded to the database in real time. [↑](#footnote-ref-1)
3. Complete descriptions of datasets for all lines can be found in Annex 1. [↑](#footnote-ref-2)
4. <https://www.nssl.noaa.gov/education/svrwx101/lightning/forecasting/> [↑](#footnote-ref-3)
5. Yet Another General Regression Neural Network <https://cran.r-project.org/web/packages/yager/yager.pdf> [↑](#footnote-ref-4)
6. <https://www.ecmwf.int/en/about/media-centre/news/2018/how-predict-lightning> [↑](#footnote-ref-5)
7. We translated variable names from Spanish to English. [↑](#footnote-ref-6)