

# DS4A — Project Report Team 60

## Characterization of Criminal Recidivism in Colombia

Coca-Castro Alejandro, Hoyos García David, Velásquez Mario, Viana Nicolás

August 3, 2020

### 1 Introduction

Recidivism is a measure of a former prisoner’s likelihood to be re-arrested, re-convicted, or returned to prison with or without a new sentence during a three-year period following the prisoner’s release. It has been used to study the performance and effectiveness of privately and publicly managed prisons.

### 2 Problem Description

The main objective is the “characterization of the recidivist population of the Penitentiary and Prison System in Colombia”. This information is key to have inputs for the formation and monitoring of the criminal policy of the Colombian State in its phases of prevision and tertiary criminalization. The objective is framed in the National Plan of Criminal Policy 2019 - 2022 of the Colombian state, which has among its expected results the reduction of recidivism. It is also important to highlight that this objective is framed in the National Plan of Criminal Policy 2019 - 2022 of the Colombian state, which has among its expected results the reduction of recidivism.

Addressing the problem of recidivism is primarily essential to prevent crime (and consequently prison overcrowding). In addition, it supports a better implementation of re-socialization treatments that characterize the convicted population susceptible to recidivism, changing the focus from punishment to semi-personalized rehabilitation programs. Even more important is the increase in public security levels.

### 3 Proposal

#### 3.1 Potential use and impact

The benefits that the Ministry of Justice could derive from the identification of patterns of recidivism are linked to the development of public policies for the prevention and re-socialization of the population confined in prisons. This is an intangible benefit that acquires greater value if it is applied, for example, in detention centers for the development of programs of counseling, training and treatment of offenders (semi-customized) or if it is used as a criteria for sentence enforcement judges to assess their decisions with a greater degree of objectivity (for example the benefits of extramural detention). That being said, there is no economic benefit, nor is there an indicator of decreased crime or recidivism that improves performance as a result of the characterization of recidivism patterns, but rather this knowledge serves as an enabler for better decision-making by the competent authorities.

In 2018, a special report in SEMANA magazine indicated that the amounts associated with the re-socialization of the prison population were around \$97,000,000,000 COP, with a maintenance base of one prisoner close to \$13,000,000 COP, while the annual investment of the Colombian state in one student was \$2,150,000 COP. Therefore, the magazine claimed that prisons were an inefficient and costly solution. Based on our characterization work, we hope that in the medium term the authorities could take better advantage of the investment amounts destined to re-socialization and consequently reduce the number of re-offenders, therefore the prison population with the amount of maintenance that it represents for the nation's finances.

## 4 Datasets sources

### 4.1 Main dataset

Our main dataset is derived from a public database curated by the Ministry of Justice and Law. The dataset (161362 registries) contains information of the post-sentenced population that has been in recidivism from 2010 to 2019. This dataset contains information on the specific crime committed, individual's legal status, date of admission to the prison, basic demographic features (age and sex), and the type and place of detention (regional, name).

We also received four additional datasets from the Ministry (222840 registries total), which contains information on the non-recidivist prison population from 2010 to 2020, with details of the crime charged, dates of capture, entry/exit from the place of confinement, legal status, age, gender, place and type of confinement.

We highlight the relevance of having information on both recidivist and non-recidivist prison populations. However, according to the theories of James Bonta and Donald Andrews (The Psychology of Criminal Conduct) and their integrative model of criminal behavior called Risk - Need - Responsibility (RNR), in our data set we do not have information on the 4 factors of greatest risk of recidivism:

- The individual's history of antisocial behavior that is characterized by the early and recurrent onset of various antisocial behaviors, with emphasis on those that result in punitive offenses.
- Antisocial personality pattern
- Anti-social rationalizations (values, beliefs, attitudes)
- The antisocial network to which the individual may have belonged (or still belong).

These factors are characteristic of the individual and determining them requires a personalized diagnosis by a specialist; also the information is not systematized or available to the public.

## 5 Data Wrangling and Data Cleaning

This step should should data preprocessing such as data aggregation, consolidation, standardisation of data types (e.g. dates, correct strings format, etc.), evaluating missing data. Quality Data produces are compulsory to generate reliable analyses for decisions, therefore it was necessary to improve the consolidated dataset. First of all, an initial evaluation of the health of the information was performed to observe the issues the data had. The following list describes better the findings:

- Non-parsed dates;

- Missing and null values in some columns, mostly related to dates of release (convicts still in prison);

MISSING OR NULL VALUES	Count
interno	0
delito	0
tentativa	0
agravado	0
calificado	0
fecha ingreso	0
fecha salida	46717
fecha captura	0
situacion juridica	0
año nacimiento	8
edad	8
genero	0
estado civil	288777
pais interno	9328
departamento	8
ciudad	0
reincidente	0
estado ingreso	52
actividades trabajo	0
actividades estudio	0
actividades ensenanza	0
nivel educativo	0
hijos menores	0
condicion expecional	334088
codigo establecimiento	0
establecimiento	0
depto establecimiento	6447
municipio establecimiento	6447
regional	0
estado	0

- Variables types not formatted properly (e.g. categorical, object);
- Sparse crimes (more than 200 different crimes and only a few committed per convict);
- Multiple classes in just one column ("Condicion Excepcional") and not all convicts have values in it since are related to race, sexual preferences, religion, etc.

After detecting what needed to be done, all the listed problems detected above were fixed. The null and missing values were handled as follows: "fecha salida" was left NaN", "estado civil" whole column was removed, "pais interno" was changed to unknown if blank, if "estado", "municipio", or "ciudad" was missing, it was acquired from departamento. Finally, as "condicion especial" had most of the observations blank, it was deduced that the minorities can be represented in a separate column to emphasize their special characteristic, and removed the previous column as most of the interns were not cataloged with a special condition.

Once the main dataset was properly wrangled (reformatted and standardized), it was exported and included in a postgresQL database on an Amazon RDS instance containing more than 160,000 entries and 30 variables from a unified dataset of 120MB. We identified inconsistencies regarding dates from the multiple original datasets and performed the necessary changes. The datatypes from the database were properly set to allow manipulations using SQL queries.

## 6 Exploratory data Analysis

Initial exploration of the data is used to answer the questions: *what happened? and the reason of these events or activities*. Thus, a basic statistical summary description was performed on each one of the variables (columns) and a first glimpse can be describe as follows:

- Total registers: 365,339
- Total initial columns: 30
- Date Range: From 2010 to 2019
- Recidivists count: 78,683
- Non-recidivists count: 173,714
- Most re offenses committed by one convict: 23
- Most popular crime: Theft
- Average age: 39 years old
- Most representative age range: 29 - 35
- Most common civil state: "Union Libre"
- Sex share: Male 89% - Female 11%
- Special conditions: 78
- Establishment with most recidivists: "Complejo Carcelario y Penitenciario Metropolitano de Bogota"

After splitting the data frame into recidivists and non-recidivist, some insights were visible:

- For the population of convicts who re-offended, the average age is 3710 years old. The average number of offenses per individual equals to 1.81.1.
- For the population of convicts who re-offended, the average age is 3912 years old. The average number of offenses per individual equals to 1.20.7.
- For the whole population of convicts, the average age is 3911 years old. The average number of offenses per individual equals to 1.50.9.

Visualizations were a way to clearly see various differences considered within the target variable. In the first example, it can be seen the comparison of age distribution (in Figure 1) between both groups as a density plot.

Secondly, a grouping method was performed to observe the distribution of regions as shown in Figure 2.

Now that there was an idea of how it was distributed between regions and ages, it was adequate to plot an inclusive demographics chart considering gender and age ranges (Figure 3).

Turning the focus into the crimes, as this variable is quite wide in classes, it was decided to plot it as a block treemap chart as a way to find their proportions, but later on the report it will

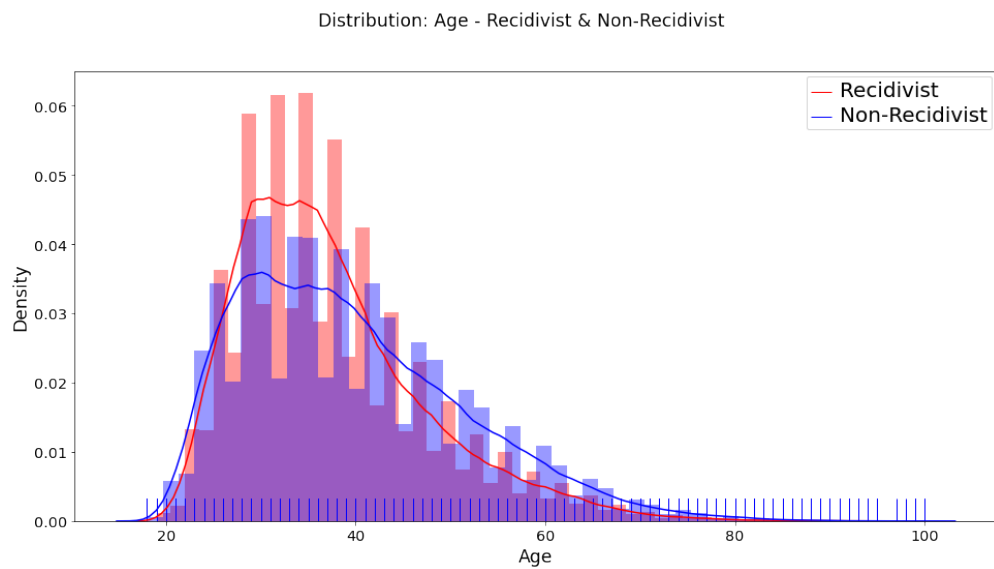


Figure 1: Age Distribution between Recidivists and Non-recidivists

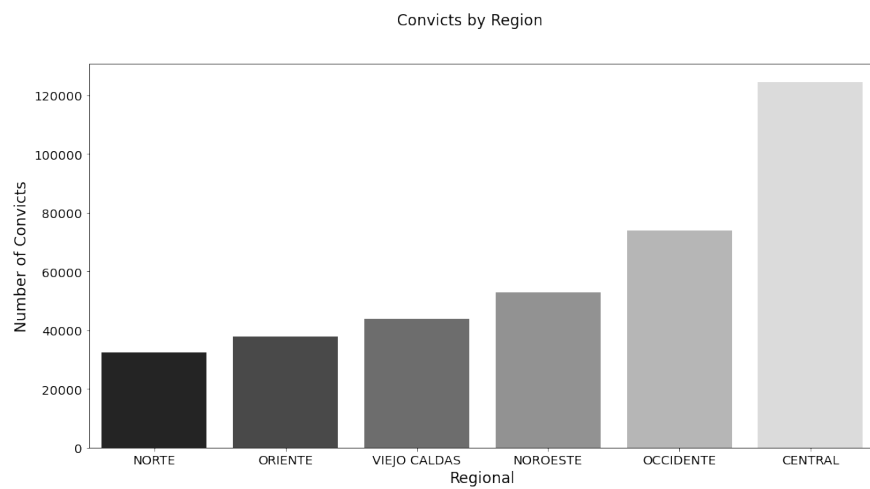


Figure 2: Number of convicts per region

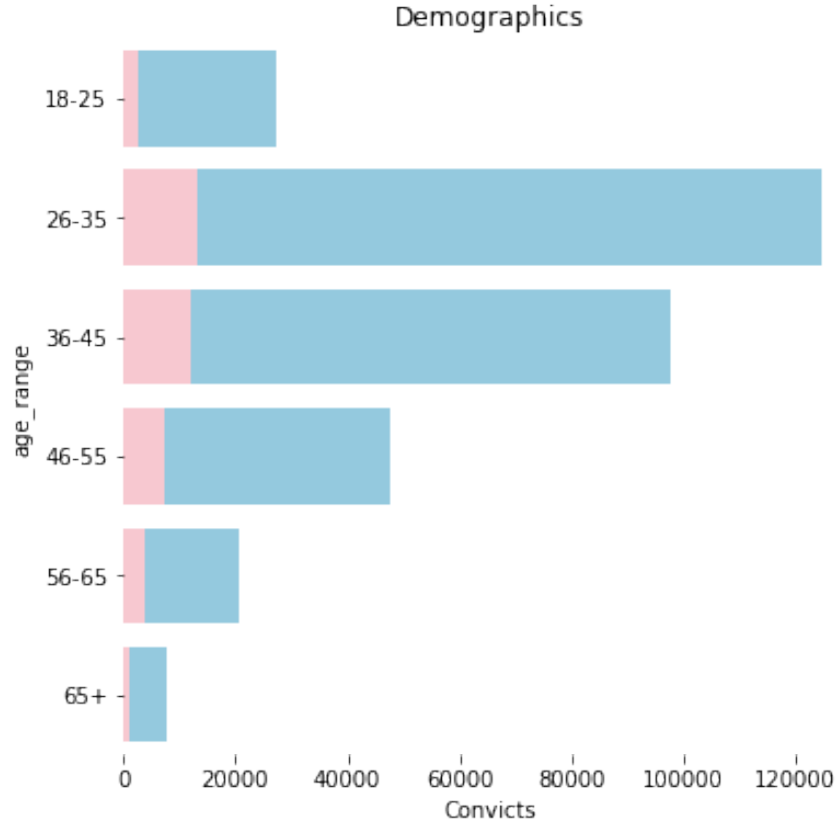


Figure 3: Demographics Distribution Chart

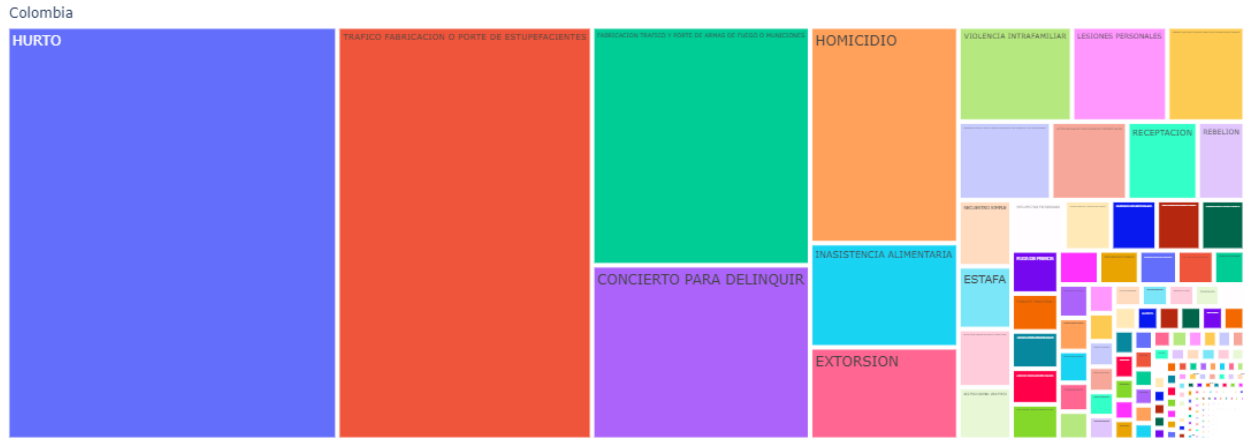


Figure 4: Crimes share treemap Chart

be explained a better way to group them and visualize them in a more comprehensive way that helped the recidivism classification groups. The figure 4 displays the first approximation.

Following crimes analysis, a word-cloud chart was depicted to get the most common words in crimes (exposed in figure 5).

An advanced aggregation was needed to obtain the number of years spent in prison of each convict, and the result is describe the figure 6.



Figure 5: Crimes words cloud

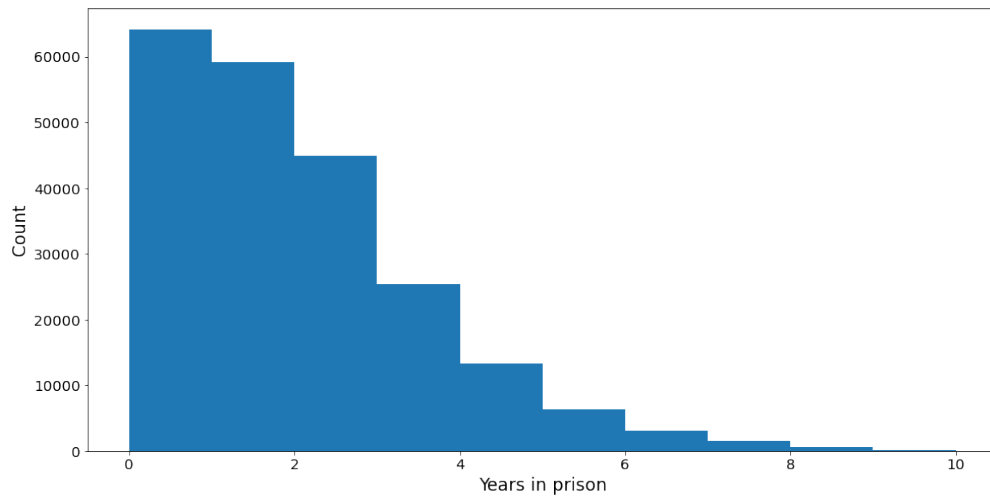


Figure 6: Time of Sentence count

Finally, a temporal analysis was made comparing both entry and exit dates for months and years behavior. The result is the following figure 7.

## 7 Feature engineering

As it was shown in the EDA, there are some features like *age* and *gender* that showed interesting results when analyzing the recidivist and no recidivist population. However, there are some other important features that can be gathered from the existing ones. Additionally, there are some features that are not easy to feed into a model the way they originally come in the data source and therefore need some adjustments and modifications:

- *Age*: The original feature that was available in the data set refers to the current age of the convict which is not accurate because it does not reflect the convict situation at the moment when he/she was processed. This variable was re-calculated using the entry date and the birth date of the convict in order to know the age back when the individual committed the crime(s).

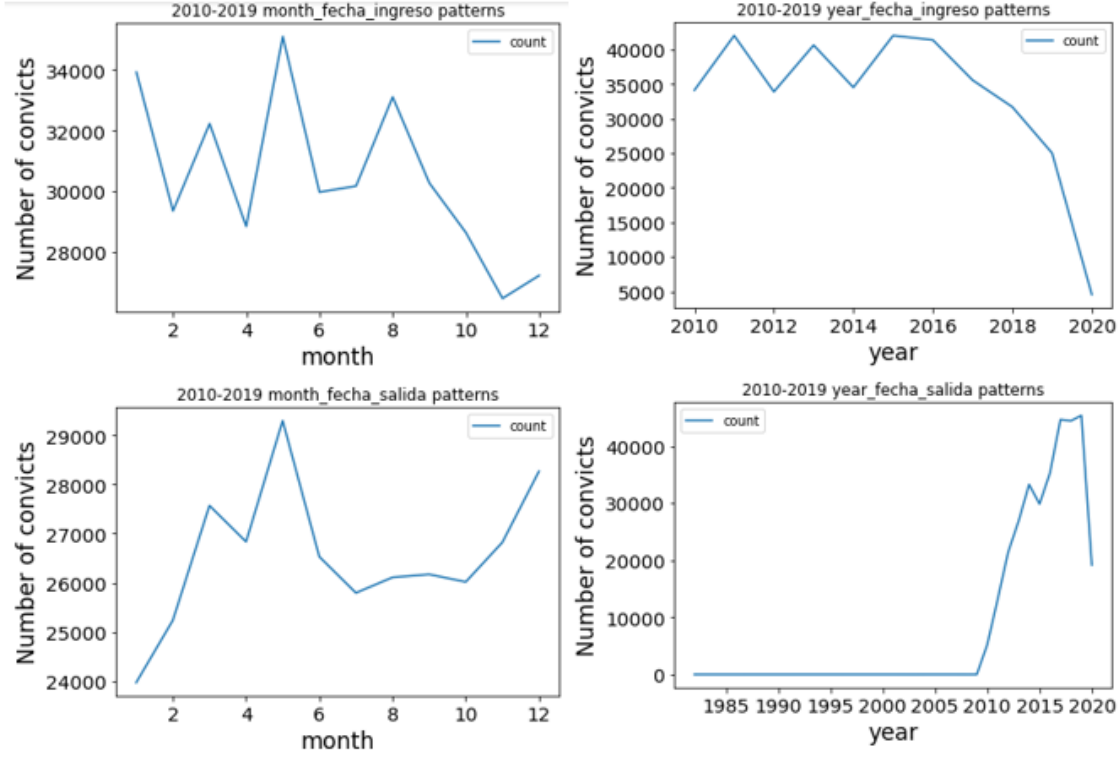


Figure 7: Temporal Analyses for month and year behaviour

- *Sentence time*: This was calculated using the entry date and release date of the convict. It is more likely for a convict who spent 2 years in prison to be recidivist than someone who were in for 40 years.
- *Special condition*: The information contained in this feature had many not standardized text categories. The information contained here was used to summarize the feature creating 6 new binary features: 'Afro', 'Indigena', 'Adulto Mayor', 'LGBTI', 'Discapacitado' and 'Madre lactante/gestante'. All these features give a better understanding about special characteristics of the individuals.

The information on its original source has a single line for each individual/entry date/crime that was committed. Since the objective is to have a prediction for each individual, the information was grouped to individual/entry date level. This grouping allowed additional feature engineering:

- *Crime count*: After grouping the information the crime count was calculated by counting the number of crimes committed by the convict on the entry date.
- *Tentativa (grouped)*: If all the crimes were labeled as 'Tentativa' then the feature will be 'Yes' otherwise 'No'.
- *Agravado (grouped)*: If any of the crimes was labeled as 'Agravado', then the feature will be 'Yes' otherwise 'No'.
- *Calificado (grouped)*: If any of the crimes was labeled as 'Calificado', then the feature will be 'Yes' otherwise 'No'.



Additionally, a deeper analysis over the specific crimes committed by the individuals was conducted:

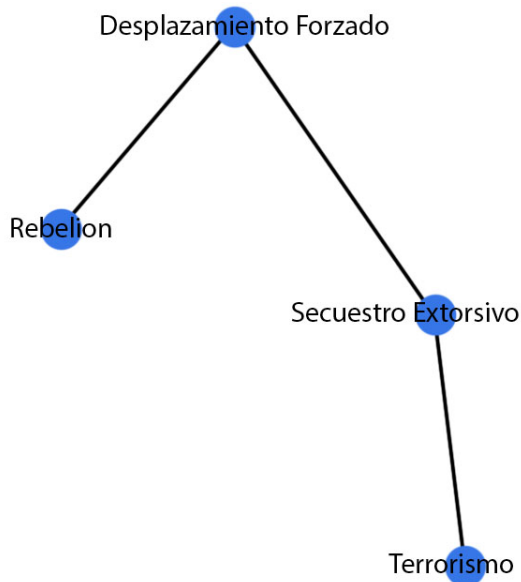
## 7.1 Clusters

Considering that our data contains more than 250 features regarding crime and that it is of interest to describe specific crime regularities, a clustering algorithm was put in place to describe crimes based on step-wise patterns to outline which specific offenses will follow up. This clustering method allowed us to reduce the amount of features and include certain crimes as part of the predictive algorithm.

This was accomplished by taking all the observations from the recidivist population and calculating for each crime the pairwise quantities in a similarity matrix  $J$  where each element describes the rate for crimes  $A$  and  $B$  as

$$\frac{A \cap B}{A \cup B}$$

In other words, it describes the similarity between two crimes as the rate of people who committed  $A$  and  $B$  divided by the amount of people who committed either of them. From this matrix, a network was constructed where each node represents a crime and each edge the relationship between both felonies weighted by the similarity index. Once a complete network was built, it was split into clusters using a minimum spanning tree algorithm based on the weights matrix, which takes every loop from the original graph and transforms it into a tree by removing the edges with the lowest weight/similarity. This algorithm works in this context since we are keeping only the relationship of crimes with the highest probability of occurrence. As expected, this unsupervised clustering algorithm grouped the crimes in a consistent manner showing clear similarities.



From the example in Figure 8 we can conclude for instance that people involved in terrorism acts have a high probability of continuing with armed conflict related crimes after their release, but have a higher probability of proceeding afterwards with extorve kidnapping.

The criminal acts that the clustering algorithm built are:

- Cluster 0: Acts involving fraud, deception, or corruption

- Cluster 1: Armed conflict and injurious acts of a sexual nature
- Cluster 2: Coercive actions, extortion, blackmail, torture, and exploitation
- Cluster 3: Acts related to arms trafficking and illegal use of military or police equipment
- Cluster 4: Acts causing harm or intending to cause harm to a person, burglary and domestic violence
- Cluster 5: Acts leading to death or intending to cause death, drugs and arms possession
- Cluster 6: Information, communication or computer-oriented crime
- Cluster 7: Acts related to the appropriation or embezzlement of public funds
- Cluster 8: Copyright infringement, counterfeit and financial crimes
- Cluster 9: Terrorism and acts against public safety and national security
- Cluster 10: Acts related to natural resources and chemical substances
- Cluster 11: Drug trafficking

## 8 Modeling

### 8.1 Significant feature selection

Once the feature engineering was finished the task was to define from all the available variables the ones that contribute the most to the objective: *Predict the probability of an convict to be non recidivist/recidivist given a set of features*. An initial logistic regression model was run using the following features:

- |                 |                       |                   |                       |
|-----------------|-----------------------|-------------------|-----------------------|
| • Age           | • Agravado            | • Underage kids   | • Disabled person     |
| • Sentence time | • Calificado          | • Education level | • Pregnant mother     |
| • Crime count   | • Gender              | • Afro            | • 13 Cluster features |
| • Region        | • Work activities     | • Indigena        |                       |
| • Sentence type | • Study activities    | • Elderly person  |                       |
| • Tentativa     | • Teaching activities | • LGBTI           |                       |

For the first model using above features, 12 of them are not significant using a 5% significance level (Figure 9):

In order to complement this analysis, an L1 regularization model (Lasso) was run using a lambda of 0.001 (Figure 11).

	coef	std err	z	P> z	[0.025	0.975]		coef	std err	z	P> z	[0.025	0.975]		coef	std err	z	P> z	[0.025	0.975]
edad	0.0127	0.001	14.531	0.000	0.011	0.014	agravado	-0.0047	0.022	-0.211	0.833	-0.048	0.039	CLUSTER_1	-1.0896	0.054	-20.172	0.000	-1.195	-0.984
sentencia	-0.1440	0.006	-25.381	0.000	-0.155	-0.133	calificado	0.2371	0.029	8.159	0.000	0.180	0.294	CLUSTER_2	-0.0762	0.035	-2.187	0.029	-0.145	-0.008
cuenta_delitos	0.1478	0.019	7.606	0.000	0.110	0.186	genero	0.2790	0.026	10.571	0.000	0.227	0.331	CLUSTER_3	-0.1162	0.055	-2.108	0.035	-0.224	-0.008
regional_CENTRAL	-0.5808	0.568	-1.023	0.306	-1.694	0.532	actividades_trabajo	0.6138	0.017	35.853	0.000	0.580	0.647	CLUSTER_4	0.5235	0.037	13.986	0.000	0.450	0.597
regional_NOROESTE	-0.6958	0.568	-1.225	0.220	-1.809	0.417	actividades_estudio	0.8379	0.019	44.017	0.000	0.801	0.875	CLUSTER_5	-0.3002	0.034	-8.909	0.000	-0.366	-0.234
regional_NORTE	-0.3854	0.568	-0.678	0.498	-1.499	0.728	actividades_enseñanza	-0.1631	0.051	-3.198	0.001	-0.263	-0.063	CLUSTER_6	-0.1854	0.101	-1.836	0.066	-0.383	0.013
regional_OCCIDENTE	-0.6641	0.568	-1.169	0.242	-1.777	0.449	hijos_menores	-0.2296	0.021	-11.113	0.000	-0.270	-0.189	CLUSTER_7	-0.7923	0.055	-14.304	0.000	-0.901	-0.684
regional_ORIENTE	-0.5831	0.568	-1.026	0.305	-1.697	0.531	nivel_educativo	-0.0193	0.006	-3.326	0.001	-0.031	-0.008	CLUSTER_8	0.0395	0.099	0.397	0.691	-0.155	0.234
regional_VIEJO CALDAS	-0.5602	0.568	-0.986	0.324	-1.674	0.553	AFRO	0.6412	0.047	13.774	0.000	0.550	0.732	CLUSTER_9	-0.8526	0.080	-10.647	0.000	-1.010	-0.696
estado_ingreso_Detencion Domiciliaria	-0.5348	0.566	-0.944	0.345	-1.645	0.575	INDIGENA	-0.0326	0.084	-0.388	0.698	-0.197	0.132	CLUSTER_10	-0.6079	0.073	-8.331	0.000	-0.751	-0.465
estado_ingreso_Espera Traslado	0.2956	0.618	0.479	0.632	-0.915	1.506	ADULTO_MAYOR	-0.0074	0.059	-0.125	0.901	-0.123	0.109	CLUSTER_11	-0.0333	0.034	-0.978	0.328	-0.100	0.033
estado_ingreso_Intramuros	-1.3751	0.566	-2.430	0.015	-2.484	-0.266	LGBTI	0.0980	0.092	1.065	0.287	-0.082	0.278	CLUSTER_12	-0.0526	0.127	-0.413	0.680	-0.302	0.197
estado_ingreso_Prision Domiciliaria	-0.8748	0.566	-1.546	0.122	-1.984	0.234	DISCAPACITADO	0.4042	0.081	4.962	0.000	0.245	0.564							
estado_ingreso_Vigilancia Electronica	-0.9641	0.567	-1.701	0.089	-2.075	0.147	MADRE_LACT_GEST	0.0485	0.176	0.276	0.782	-0.296	0.393							
tentativa	0.1116	0.937	0.026	0.802	0.839	0.184	CLUSTER_0	0.0991	0.051	1.949	0.051	-0.001	0.199							

Figure 9: Initial logistic regression with all variables.

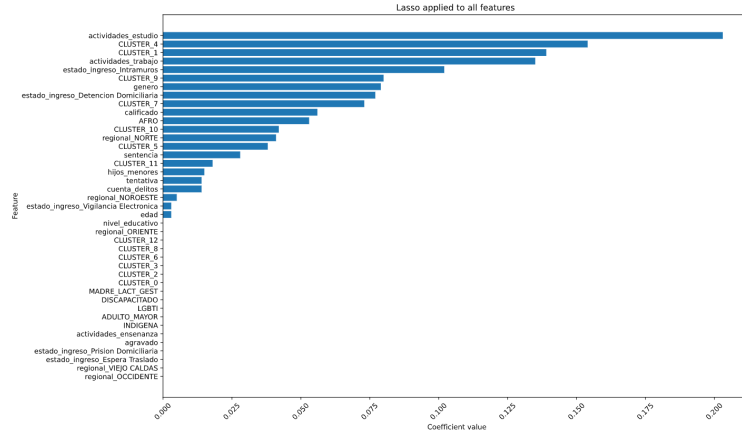
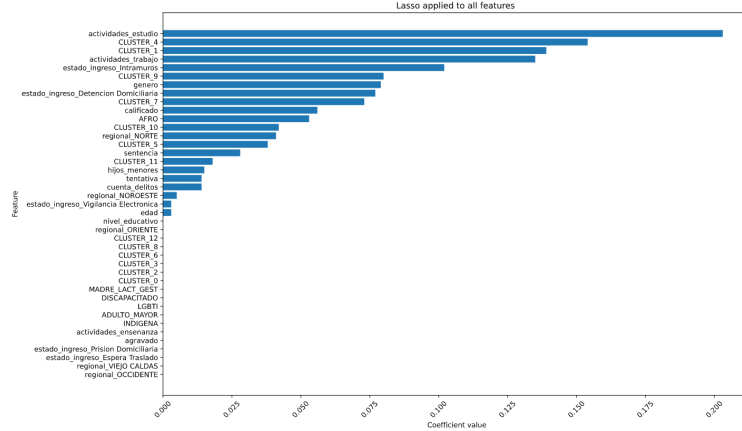


Figure 10: Lasso coefficients (lambda = 0.001).



we see that among all the possible values for this variable, 'Intramuros' covers around 60 % of the population and would help to distinguish if the convict is actually inside a prison or not. Additionally many of these *Sentence type* features have a 0.0 coefficient in the lasso model. Considering this, only the dummy feature 'estado-ingreso-intramuros' will be included in further models.

Looking at the *Region* feature we see that almost all of them got 0.0 coefficient in the lasso regression. After comparing baseline models AUC with and without the *Region* feature it was found that it does not increase the predictive power of them, also these feature showed to be non significant in the initial logistic regression. In order to keep the models simple and reduce their

Many of the coefficients of the regularization turn 0. Many of these match the non-significant variables found in the initial logistic regression model. Having both results into account the features *Indigena* , *LGBTI* , *Pregnant mother* and *Elderly person* will not be included in further models for being non-significant and having 0.0 coefficient according to the logit model and lasso values, respectively.

Even though some of the *Sentence type* dummy variables are significant,

dimensionality this features will not be included either in next models.

Finally, the clusters of crimes 1, 2, 3, 4, 5, 7, 9 and 10 showed to be significant. All the other clusters will not be used on the subsequent modeling. Similarly only the significant interactions will be included: *Cluster2-Agravado*, *Cluster4-Tentativa*, *Cluster5-Tentativa* and *Cluster5-Agravado*.

This significance and L1 regularization analysis helped us to reduce the number of features to be used in future models from 42 to 25 without affecting the predicting power of the models:

- Age
- Sentence time
- Crime count
- Sentence type Intramuros
- Tentativa
- Calificado
- Gender
- Work activities
- Study activities
- Teaching activities
- Underage kids
- Education level
- Afro
- 8 Cluster features
- 4 Interaction clusters

## 8.2 Class balancing

Since the classes of our lable (reincidente / no reincidente) are not balanced (30-70), the majority class ('no reincidente') was downsampled by randomly selecting observations of that class. The resulting dataset after the balancing process had 46.298 observations for each class and 92.596 observations in total.

## 8.3 Model selection

Once the key features were selected 4 different models were run. Each model was tuned (if possible) and cross-validated to have better comparison between models.

### 8.3.1 Logistic Regression

Using 20 fold cross-validation, multiple logit models were run. For each model the ROC and AUC were calculated and then averaged to obtain a final AUC of 0.698.

### 8.3.2 Random Forest Classifier

The random forest model was first tuned fitting multiple models varying on each the max depth for the trees of the forest. For each model the AUC was calculated in order to know which is the better configuration of this model given the available data. The tuning process shows that the best depth is 16 in this case. Using 20 fold cross-validation, multiple random forest models were run. For each model the ROC and AUC were calculated and then averaged to obtain a final AUC of 0.734. From this calibrated random forest model it is possible to obtain a quantification of the importance of each feature used in the trees in the task of predicting the non-recidivism/recidivism. Figure 13 shows the importance of each one of the features.

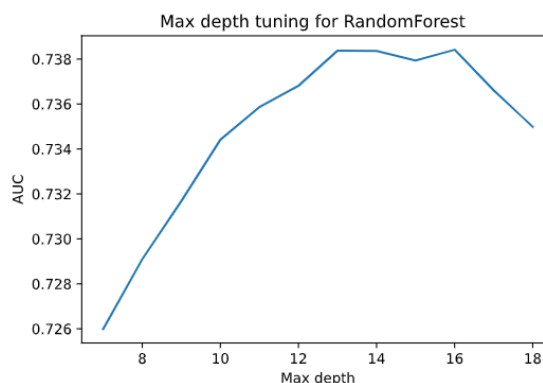


Figure 12: Tuning for RF model.

It is clear that the age of the convict is the most important factor for this model, followed by the sentence time and if the convict had study activities in prison or not. It is also interesting

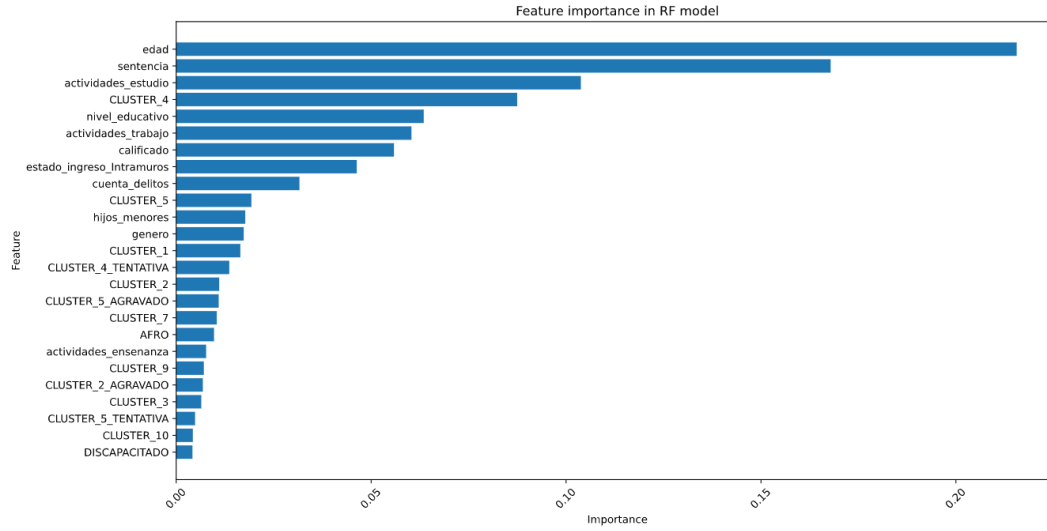


Figure 13: Feature importance in the random forest model.

that crimes in the cluster 4 (theft and violent crimes) significant when predicting recidivism. The education level and work activities in prison are also among the top 6 main features for the random forest model meaning that the intellectual/work activities may have an impact in the convicts behaviour.

### 8.3.3 Gradient Boosting Classifier

The gradient boosting model was tuned by exploring different combinations of learning rate and max depth and calculating for each combination the model AUC.

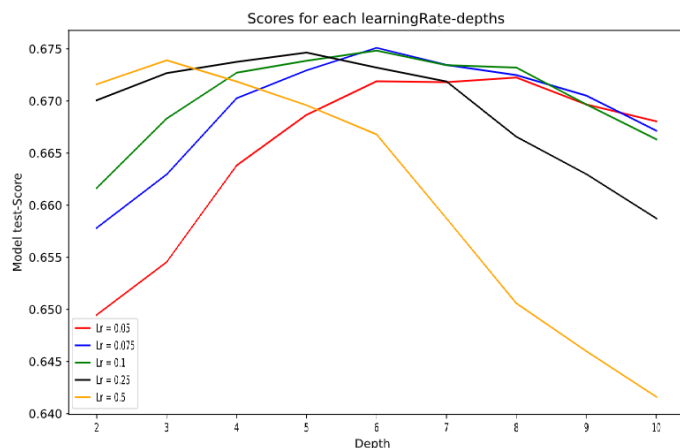


Figure 14 shows the test score for each learning rate tried. In this case, the best parameters were learning rate of 0.075 and max depth 6.

Using these results a cross-validated model was run obtaining an AUC of 0,737.

### 8.3.4 XGB Classifier

Figure 14: Max depth and learning rate for GB.

The final model tried was x-Gradient Boosting classifier. This time the tuned parameters were the learning rate, the min split loss and the maximum depth of the trees. The parameters combination with the maximum AUC where learning rate of 0.1, min split loss of 0.3 and maximum depth of 6. This combination of parameters led to an AUC of 0.737. Similarly to the random forest classifier, with XGB it is possible to obtain for each feature used in the model its weight and gain.

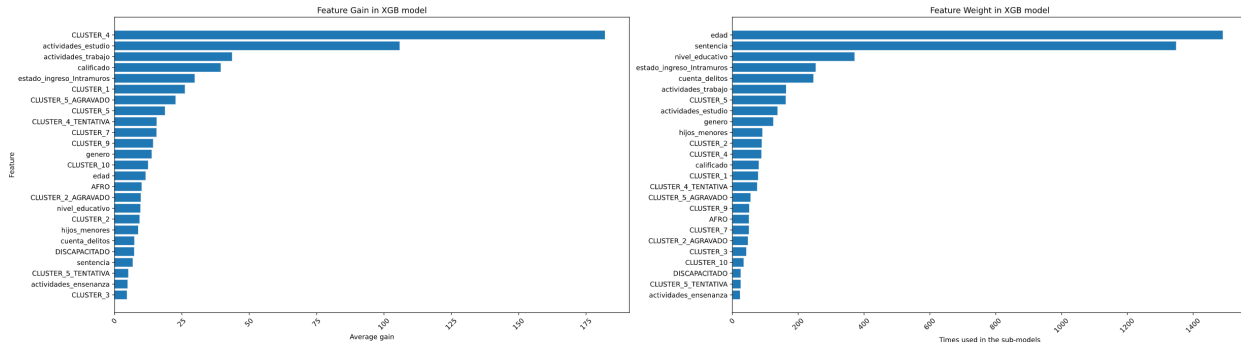


Figure 15: Features gain and weigh in XGB model.

Figure 15 shows the average gain and weight of the features used to fit the XGB model. The feature weight indicates how many times the sub-models (trees) coined in the XGB assemble used the feature to split the data while the average gain gives an idea of the average contribution of the feature to successfully splitting the model when it was used. These two metrics help to understand how often each feature is used and how much it contributes to make accurate predictions. In this case, age and Sentence length are by far the most used features to classify the recidivist/non recidivist individuals. Other features like Education level and Intramuros state are also frequently used to predict the probability of recidivism. Looking at the average gain, the cluster number 4 is the feature that in average contributes the most when is used to make accurate splits of the data.

## 8.4 Model comparison

Figure 16 shows the different ROC for each model that was tested. As it was shown before, the Logistic regression baseline model had an AUC of 0.698. After trying tuning different models the best AUC obtained was the X-Gradient Boosting (0.737).

All the models that were fitted are very useful not only to predict the recidivism probability but also to understand how the different features available contribute independently to the objective of these project.

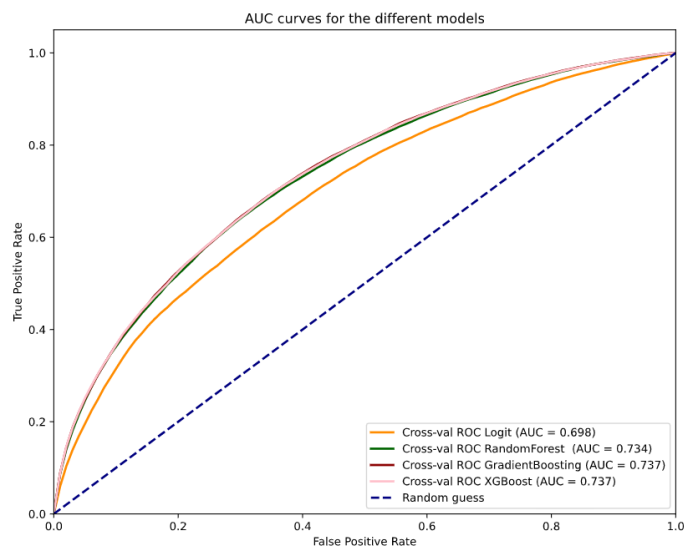


Figure 16: ROC for each model tested.

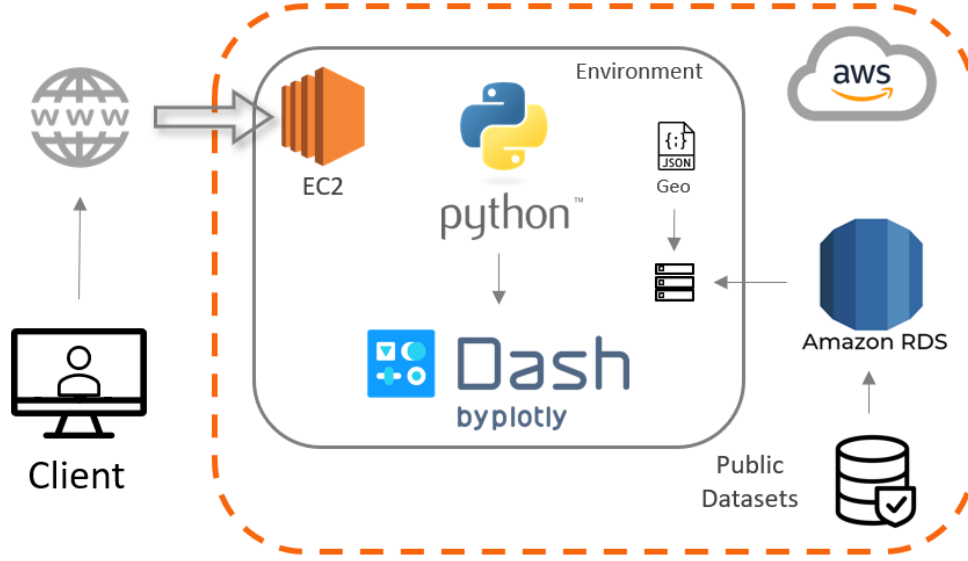


Figure 17: Architecture

## 9 Application Overview

### 9.1 Users

Government officials, private sector, non-profit organizations and/or civil society making decisions for recidivism and crimes in Colombia and reporting, control and/or management this information in their region of interest.

### 9.2 Architecture

Figure 17 shows the architecture of the proposed solution including the main elements of the application at component level and its connections at high level (see deployment diagram). Additionally, it shows the application elements used for the Front- and Back-End. The figure also indicates the key technologies used and hosted on AWS cloud i.e. Python, Dash and libraries.

The key AWS components used are as follows:

1. A machine hosting the App (Elastic Compute Cloud - EC2);
2. A Database (Relational Database Service -RDS);
3. A Security group for these services;
4. A remote GitHub repository. It contains the source code and documentation.

Architecture Scheme in Figure 17

### 9.3 Front End Design



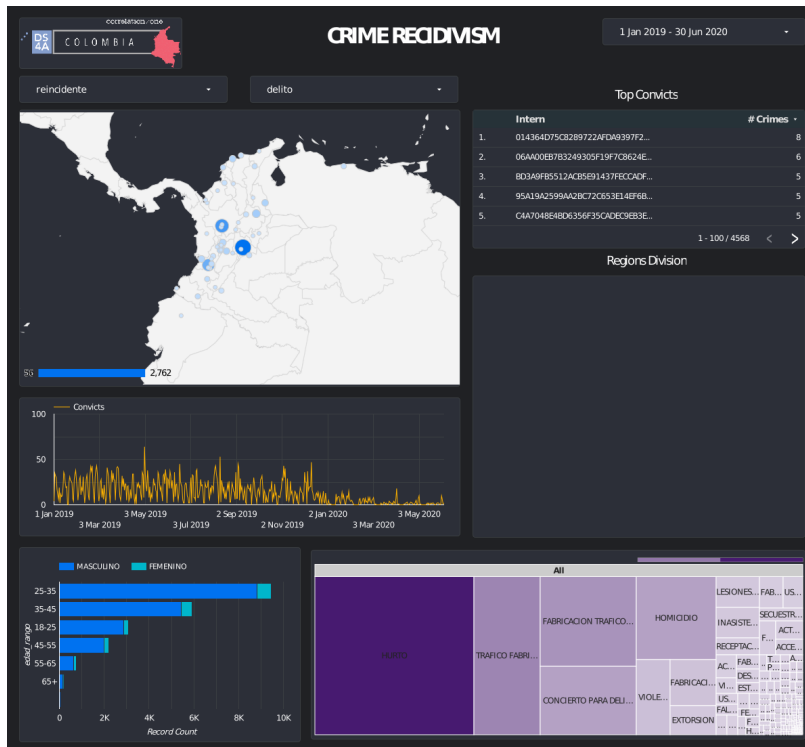


Figure 18: Front End Simulation

Overall, the proposed dashboard provides dynamic components to retrieve recidivism-related information in Colombia.

An initial draft of the solution was designed using Google's visualization tool Data Studio. This allowed to have a clearer vision of what wanted to be created. The Dashboard Mockup can be found in Figure 18

The final version of the dashboard was build under 3 main sections:

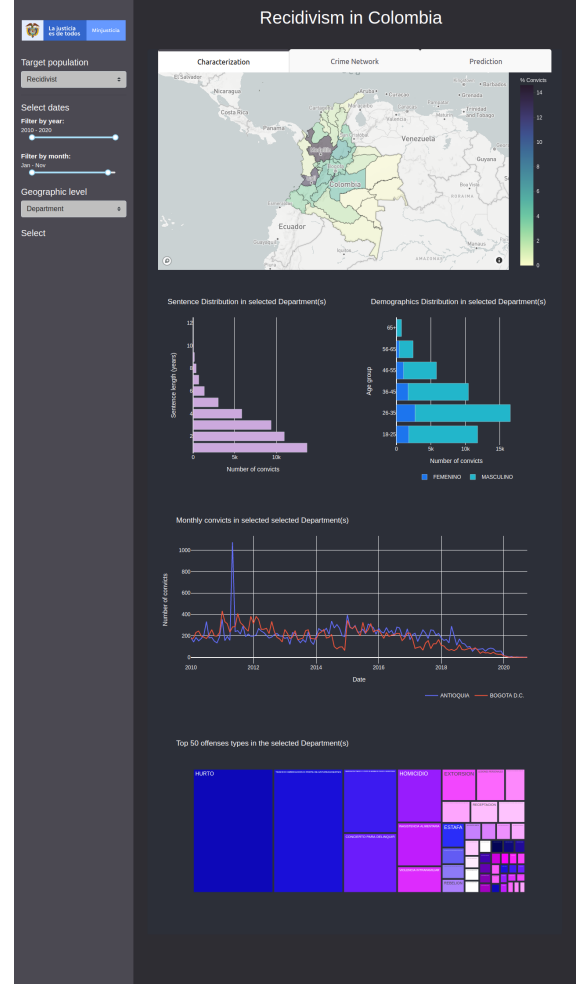
### 9.3.1 Characterization

The user will be able to geographically visualize the % con convicts on each department of the country. It also has a yer/month and target population filters that will improve the data visualisation.

This section additionally displays interesting plots about the convict population that is targeted through the section filters.

These plots show the gender distribution as well as convict demographics and sentence length distributions. Another interesting graph that is found in this section is the top 50 offenses types committed in the selected department.

This section provides useful information that will allow the user to better understand the characteristics of the recidivist and non recidivist population across the country.



### 9.3.2 Crime Network

This section (figure 20) allows the user to explore the clusters that were created during the feature engineering of this project.

Selecting a cluster from the drop down menu, will show the user the network of the crimes that are linked together. The color of the link indicates how strong is the relationship between both crimes.

### 9.3.3 Prediction

The prediction (figure 21) tab was built to give the user predictions of the probability of a convict to be recidivist/non recidivist.

After selecting the convict characteristics (age, gender, sentence length, etc) through the sliders, the probability will be calculated and displayed. This section also contains a graph that displays the importance of each feature that is used in the model to make the predictions.

Figure 19: Exploratory component

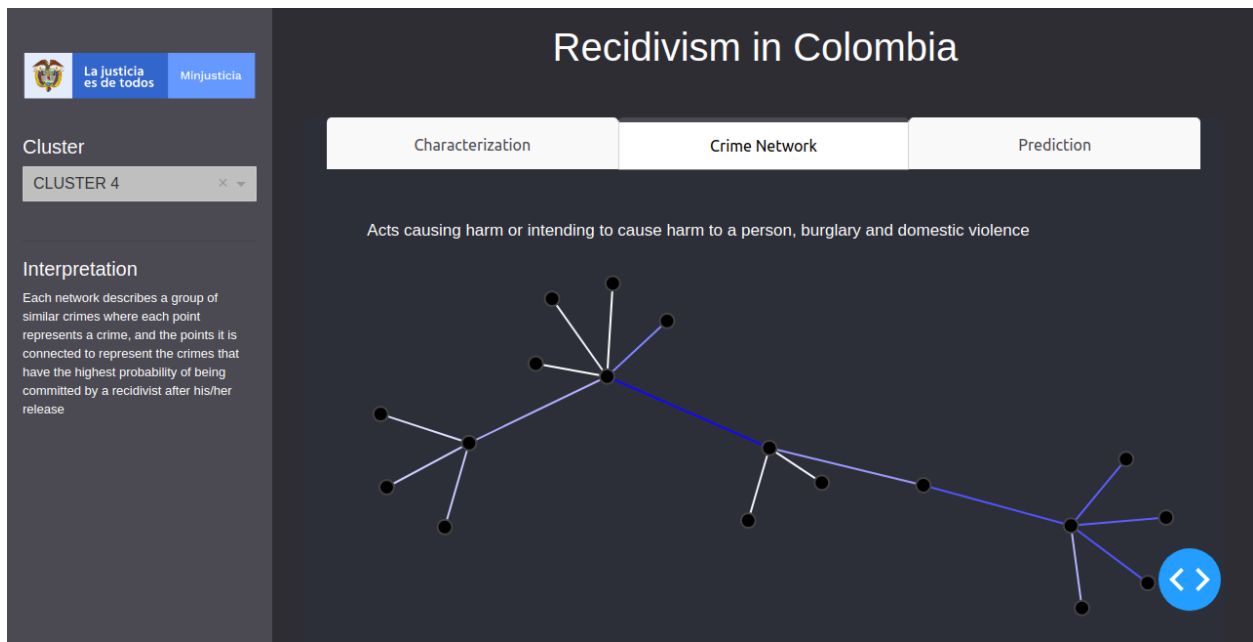


Figure 20: Cluster component

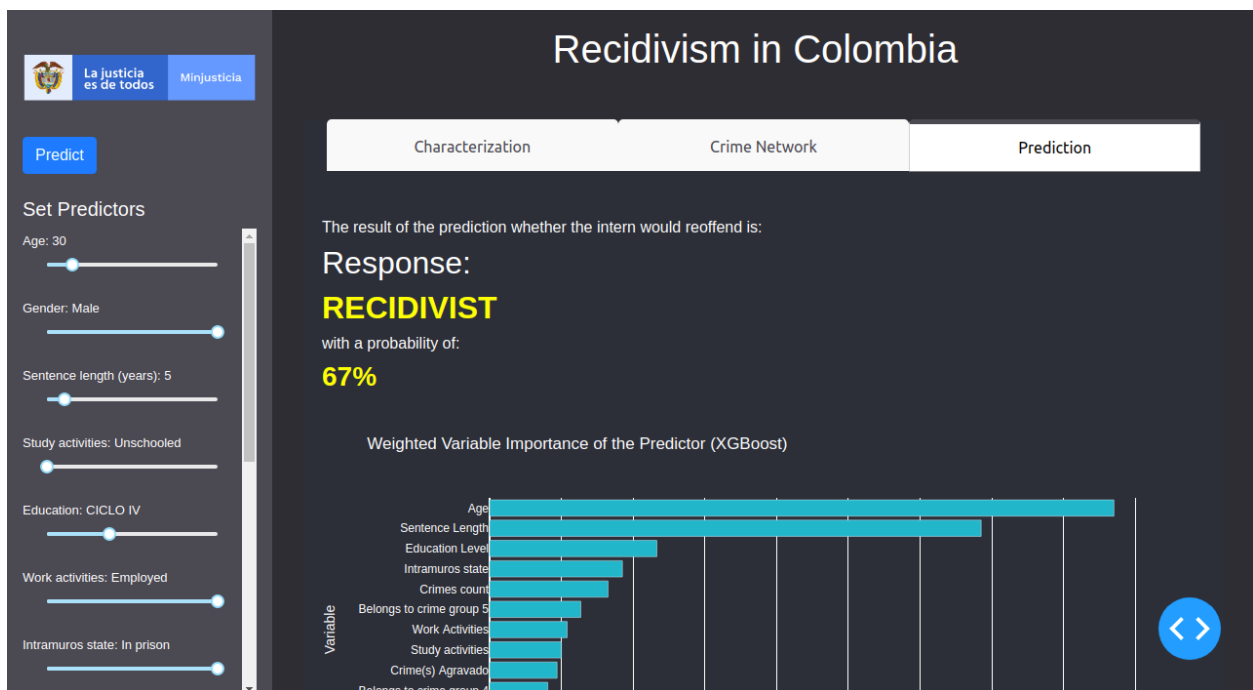


Figure 21: Prediction component

## 10 Conclusions

The initial expectations from the project were to reach the classification of people deprived of liberty in groups that allow the generation of re socialization policies.

In order to try to fulfill that statement, it was applied the whole data science process over the information provided without knowing what would be the final outcome. Some initial takeaways were drawn from exploratory analysis, which derived some patterns about recidivists, such as younger male inmates, between an age range of 29 to 35, are more prone to re offend than other groups. Another finding, was the fact that region distribution was not significant in regards to determining a criminal behavior, though more populated areas like main department capitals show peaks in the amount of crimes made by re offenders at seasonal periods such as presidential elections or holidays.

The highlights from this project, were the results of the predictive modelling stage, where it was obtained the importances of the variables that match a recidivist behavior. For example: **Age, Sentence Time, Sexual War related crimes (cluster 4), total number of crimes committed**, amongst others.

The previous discoveries couldn't be reached without the deep techniques applied during the feature engineering process, which was one of the key milestones reached since gave us the right path to follow in terms of quality, simplicity and variability of the predictors.

Afterwards, the construction and evaluation of predictive models in an incremental way, going from a simple Logistic Regression to a Random Forest and finally an ensemble using XGBoost, allowed us to leverage the advantages from each one of them. for instance, Logit provided transparency and robustness, whereas flexibility and processing optimization were obtained from the most complex models. These approach proved to be effective filtering the noise (statistically insignificant variables) from the right signal (most important features) and consequently, generalizing properly the data, and right-fitting the model into new observations.

### 10.1 Next Steps

Future work ideas are a way to let the project have continuity in the short, middle and long term. This can be done throughout the adequate handover of the project and all its resources to the technical and executive stakeholders from the ministry of Justice, assuring that everything is understandable and sustainable.

- The first work to continue with the improvement of the solution, would be the implementation of a data stream module that pulls new data from the Ministry of Justice repositories, and feeds the AWS RDS database in a frequently manner so it is up-to-date with new convicts' information. This can be done by setting up an API REST through a web service that once a day/week/month uploads the new registers.
- The Layout and charts were decided by all the team according to the importance and quality of information they provide, but it might have more sense to be discussed as well with the product owner or main stakeholder in the ministry whether this is relevant for their purposes or not and remake accordingly (Such as the translation of the text from English to Spanish).

- The predictive function was modeled based on the data provided up until the beginning of 2020. Therefore, in order to keep calibrated the accuracy and work properly, the model needs to be retrained every month or quarter with the new acquired data.
- Finally, one of the foremost steps in the long term is to truly test the hypothesis of the factors that influence the most in recidivism in the convicts. Hence, further studies and scientific experiments should be held so as to follow up the re socialization programs based on the findings of this project.

## 11 Team members

- Alejandro Coca-Castro (King's College London):
- Adolfo David Hoyos García (BPP Business School London):
- Mario Andrés Velásquez Angel (Universidad de los Andes):
- Nicolás Viana (Affiliation):