**Analysis of patterns and anomalies in judgments against the State handled by the National Agency for Legal Defense**

Team #12

Brayan Andru Montenegro Embus, Juan Esteban López de mesa, Cristian Camilo Lopez Martinez, Valeria García Múnera, Camilo Andrés Serrano Velandia.

**Overview of industry, business, or problem**

Through the Law 1444 of 2011 year, the structure of Colombian National Public Administration was modified and the National Agency for Legal Defense of the State (ANDJE, by its Spanish acronym) was created as a special Administrative Unit. The Agency is a decentralized entity of the national order, with legal status, administrative and financial autonomy and its own assets and it is attached to Colombian Ministry of Justice and Law. Through the Decree of Law 4085 of 2011, modified as pertinent by Decrees 915 of 2017, 1698 of 2019 and 2269 of 2019, the objectives and structure of the Agency were established, whose approach is oriented to the design of strategies, plans, and actions aimed at complying with the policies of Legal Defense of the Nation and the State; formulate, evaluate and disseminate policies for the prevention of unlawful conduct, unlawful damage and the direction, coordination and execution of actions that ensure an adequate defense of the litigious interests of the Nation. With the Agency, the Government seeks to unify the defense and information policies of the State in cases of strategic interest.

**Define the specific problem that should be solved**

The Single State Litigation Management and Information System (eKogui), stores fundamental fields on arbitration awards, such as: causes, entity, values, people linked to a process, among others. The problem to be solved is to create a tool based on data analytics, taking into account the historical information module of eKogui, which allows identifying and visualizing:

1. The arbitrators and the meaning of the verdict (favorable or unfavorable) for the Nation

2. Identify the arbitration issues of the "resolve" delivered by the arbitrators where an entity of the National Order is linked.

3. Atypical data in the sense of verdicts.

**Why does this problem matter?**

Currently, the eKogui system has more than 500 awards from 2015 to date where a state entity and/or the ANDJE is intervening in National Arbitration. With this tool, it would be possible to analyze and identify the sense of the awards as well as deviations or errors in the litigation activity of the State.

**Business impact**

Recording the correct information can generate improvement actions to prevent the anti-juridical damage to the State's public entities. In addition, detecting anomalies in judicial proceedings would allow the Agency to intervene promptly.

**First glance at the Datasets**

The data sets provided by the ANDJE will be queries to the Single State Litigation Management and Information System (eKogui), the dataset contains the fields:

1. Basic information on arbitration awards: includes information to uniquely identify a judicial process (Id, Jurisdiction, Action, Claims, Status of the process, if it is interinstitutional, etc.).
2. Arbitrators: convening entity, convened lawyer, name of the center, type of arbitration, ID.
3. Causes: information related to the causes of the processes: (Id, id\_causa, cause, active/inactive status).
4. Events: based on the events of the processes (Id, facts (long text)).
5. Proceedings: The record of actions on the award (id, entry, evidence, termination)
6. Counterparts: information of the defendants, plaintiffs of the processes.
7. Sentencing Values: economic values of the convictions for some processes completed in the system.
8. Economic sector (name\_ua\_sector, ID\_sector\_Ekogui)
9. Quality in which the entity acts (summoned or convener)
10. Claim/amount determined
11. Arbitration type (arbitration\_type)

The following table shows the summary of the most important features of the dasated as well as the relevant topics related with them:

| **Datasets** | **Source** | **Description** | **Topics** | **Related Topics** |
| --- | --- | --- | --- | --- |
| REPORTE\_LAUDOS\_20220315 | <https://www.dropbox.com/sh/w5fygc3f27d2sjp/AABy5McRZUgQrWbfEsJ_LHk1a?dl=0> | * From October 2007 to February 2022. * 273 arbitration awards. Number of variables: 141, between categorical and numerical. * Missing data designed as blank spaces. * Require the use of a data dictionary to translate the name of columns and categories. * Related entities, facts, place of the facts, cause, impact, selection mechanism of the arbitrators, name of the arbitrator, last action, meaning of the award, date of the arbitration award, value of the sentence against, value of the sentence in favor, type of sentence, costs in favor or against the entity, arbitration award result * Dataset includes sub-variables (secondary questions of the survey) whose values depend on the answers to primary questions. In consequence, there are missing values in all the secondary variables, and according to the specific analysis some of them may be removed. | 1. Lawsuits against state entities 2. Meaning of the arbitral award 3. Value of the judgment against or in favor of the entity | 1. Counterpart, summoned, place of events, cause, impact, general information of the arbitrator, information of the summoner, information of the summoned, last action 2. Meaning of the arbitral award, date of the award 3. Value of the sentence against, value of the sentence in favor, value set for costs, costs in favor or against the entity, if the costs must be paid by the nation |

One advantage of this dataset is that it provides important information about the arbitration awards against the state. One disadvantage of this dataset is that it contains only 273 rows, which is a very small sample size for projections. Another complication is that it will take many years to collect enough data to run efficient machine learning models.

**Methods and Models**

**Visualizations**

Review by the arbitrator how the claims are for and against and by sector and in general each type of variable as it is distributed in the sense of the award. Validate which entities were demands and classify them according to the amount. Validate the most repetitive words according to the arbitrator, type of contract, result of the ruling, sectors, higher and lower amounts.

* Bar charts (univariate), stacked bar charts (multivariate) to explore distribution of the data (arbitrator, entity, sector) by dividing it by award result and type of conviction
* Wordclouds to find the most common words per demand.
* Scatterplots and heatmaps to visualize correlations between entities, summoners y summoned with the economic values of the awards
* Single and multi-line charts to track demand amounts over time.
* Choropleth maps to visualize geographical pre-defined areas colored and patterned in proportion to the number of demands, demand amounts, results of the arbitration awards by departments, municipalities, and citys.
* Histograms to analyze the distribution of demand amounts and check if they are normaly distributed, data is skewed and presence of outliers.
* Boxplots to display the summary of behavioral variables (minimum, quartiles, median, maximum and presence of outliers).

**Which libraries to use**

For initial phases of the project, pandas and pandas profiling allow an effective data exploration. For more detailed visualizations, scatterplots, boxplots, line charts and histograms are some of the Matplotlib functionalities. For interactive reports, Seaborn, Bokeh and Altair could be good choices.

When predicting, scikit-learn and keras may be useful to create different models and compare them according to some error metrics within a test set.

**Models**

Another key component of our project is determining the association between entities, facts, place of the facts, cause, arbitrator and the meaning of the award. This project includes the use of:

● Statistical tests to find significance differences between subgroups.

● Feature engineering: Normalization, Scaling, One-Hot Encoding of variables before applying ML techniques.

● Unsupervised Machine Learning Models: Dimensionality Reduction, K-Means model to identify clusters or groups of customer profiles.

Other models/techniques to explore:

● Supervised ML for classification problems: Neural Networks, Decision Trees for classification, Logistic Regression, KNN, SVM, Naive Bayesian methods.

● Supervised ML for regression problems: Neural Networks, Decision Trees for regression, Linear (Multi) Regression.

● Unsupervised ML for classification problems: PCA, t-SNE, k-Means.

● Feature Engineering: Encoding techniques.

● Natural Language Processing techniques.

● Deep Learning Models.

**Interface**

The final front-end product will feature two landing pages: an Analytics page with visualizations of the historical data, and a Prediction page, where according to relevant inputs about the demand (arbitrator, entity, money amount, sector, etc) the result of the award will be predicted.

The interface will allow for interactive visualizations of the historical data, so that users can click on a particular sector/arbitrator/year, and see the results for that variable or the summary of all outcomes.

