# Classification model to detect patients with life at risks from the "Right of request" dataset in the Colombian Health Care system

Mario Rugeles Pérez

November 2018

## 1 Project Overview

The "right of request" is a legal mechanism included in the Colombian Constitution of 1991 designed to allow efficient communication between a citizen and the government [2]. This mechanism is one of the most fundamental democratic tools in Colombia as it allow all citizens to ensure the government is fulfilling his obligations with the people. Once a "right of request" is raised by a citizen, the government has from 10 to 30 days to resolve the case.

This in practice, means that the government needs to ensure it has the procedures and infrastructure to resolve a right of request as soon as possible.

## 2 Problem Statement

One of the basics rights of every citizen in Colombia is the access to the health care system. As part of the duties for ensuring this mission, The government of Colombia has been creating online resources to allow citizens to raise complaints or ask for information in all public entities, included The National Health Superintendent (Supersalud) [6].

Because the access to internet has been increasing in the recent years, it's expected that the volume of "right of request" will increase over time too. Supersalud receives thousands of right of request per month and part of its task is to decide which cases have more priority as many of these cases are about citizens whose life might be at risk.

Deciding when a person's life is at risk may require experts in the field of health to make the right assessment. But the volume of cases can eventually exceed experts's capacity to respond quickly with the right decision.

A data set from the "right of request" from Supersalud is available to the public through the Government's Open Data portal [1]. This data set ([3], [4], [5]) contains information related with the patient and a feature indicating if the patient's life is at risk. This data set is suitable for a classification model than

can predict if new patients rising new "right of request" have their life at risk automatically saving time for both the government and the patient.

Every record provides information about patient condition, demographic information and the health company provider that supplies health care services.

# 3 Metrics

The model must have a high recall metrics because it needs to avoid false negatives, as we don't want people with life at risk as not in risk. A initial beta of 2 will be used in the F-beta score metrics.

# 4 Data Exploration / Exploratory Visualization

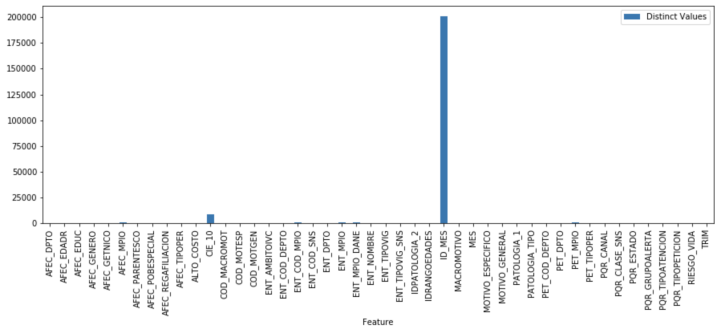
The data set being used for this project is available from Colombia Government's Open Data project [1]. Supersalud also provided a document with the feature's description that will be also used here (MANUAL BASE DE DATOS MATRIZ DATOS ABIERTOS.DOCX).

All features in the data set are categorical. They provide information related with the citizen raising the right of request. The data set provides information about the patient, like location, illness related info as well data from the Health Care company the patient is inscribed.

All the data is stored per year, new records are added is released every year in a new data set. With every new data set release every year, it's possible that few new features are added, also is expected that some can be deprecated, so a first step is to join the data sets used in a single one containing all common features.

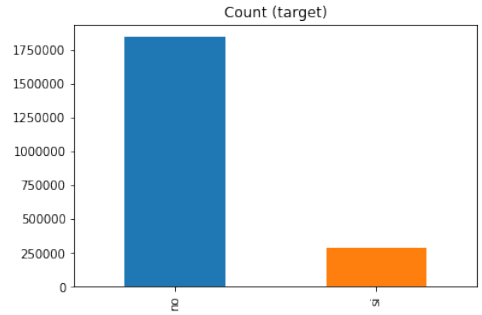
A first visualization shows high imbalanced categorical features with high cardinality (Figure 1).

Figure 1: High cardinality



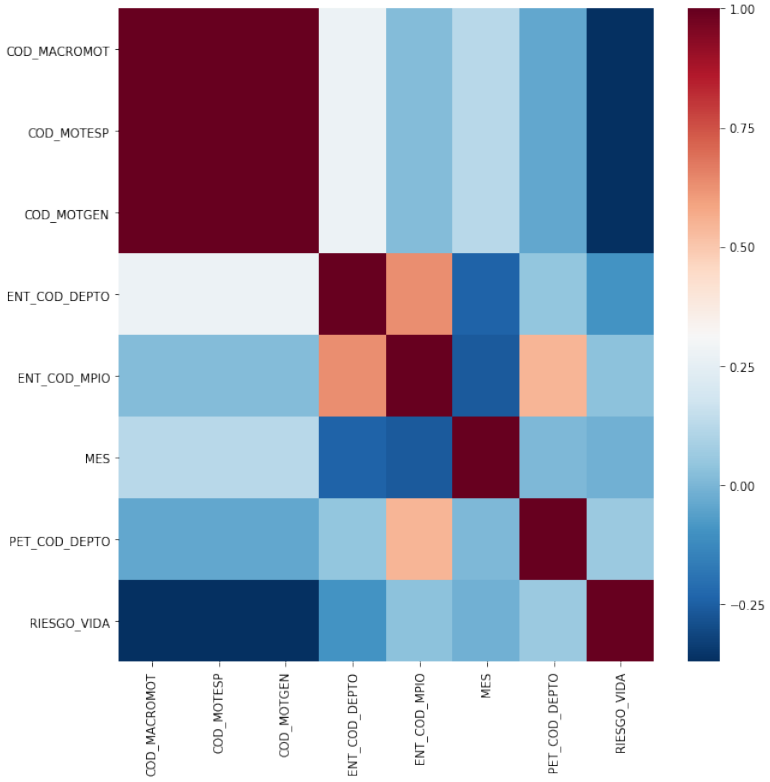

The labels also show a high uneven distribution (Figure 2). Re-sampling will be taken into account in the dataset prepossessing.

Figure 2: Unbalanced target




A initial analysis of correlations over the target column RIESGO_VIDA shows very poor correlation with other features. It's very likely to be due to the data sparsity (Figure 3).

Figure 3: Low correlations with target feature RIESGO_VIDA



# 5 Algorithms and Techniques

The features in the dataset are exclusively categorical, there are a total of 43 features, so technique of one hot encoding will not be used in this project as it will increment extensively the number of features in the dataset (Feature CIE_10 for example, has about 9000 values). Some records in some features have the same meaning despite having different values.

A initial cleaning over the dataset will be performed to address incorrect values in the features. Also, according to the official documentation some features lack of statistical meaning, so it will removed.

Mean Encoding will be applied as a solution to the high cardinality issue, also values transformation will be applied to normalize the dataset to address the low correlation weight with the target variable.

Given the extremely low correlation with the target variable, and the fact that there are considerably many features, the model will require to rely on a combination of boosting learners and stacking methods to push the score as high as possible. Over and Under sampling will also be taken into account.

# 6 Benchmark

A Logistic Regression model will be use as a base model to compare with the final model.

# 7 Data Preprocessing

## 7.1 Removing fields

According to the oficial documentation, fields "IDRANGOEDADES", "ID_MES" and "PQR_GRUPOALERTA" have not statistical use, so they are removed from the dataset.

On the other hand, the feature "PQR_ESTADO" has a significant statistical value that may bias the model: Once a right of request enters the system, it goes through a series of states before the case is closed (That state is stored at "PQR_ESTADO"). Historically, patients with life at risk can have a tendency to have a closed state as they may have priority over other cases, so including "PQR_ESTADO" will make the model to make predictions over a feature that will not be available when introducing a new right of request (When a new right of request enters the system it will have a default state that is very unlikely to have the final state from the original data set).

## 7.2 Special cases: AFEC_DPTO, PET_DEPTO

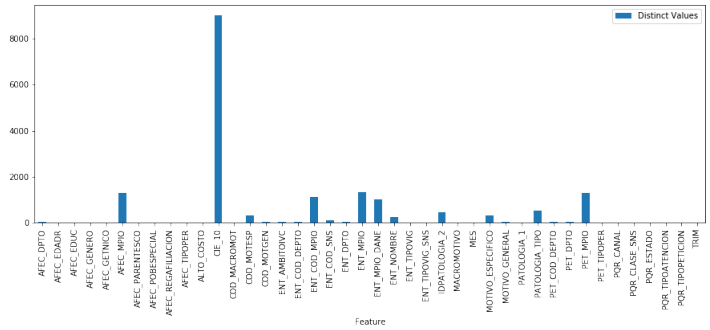
Having a initial look to the features, features AFEC_DPTO and PET_DEPTO show some repeated states with different names, and given that there are rela-

tively few states, it's possible to manually fix these values to avoid records with different names but refering to the same state.

Some feature's values are written in different ways, for example, 'ARCHIPIELAGO DE SAN ANDRES, PROVIDENCIA Y SANTA CATALINA', 'SAN ANDRES' and 'SAN ANDRÉS' refer to the same state. Same for 'BOGOTA D.C' and 'BOGOTA D.C.'

After this initial cleaning process some features start to show more statistical meaning than the previous dataset (Figure 4).

Figure 4: Some features gain statistical meaning



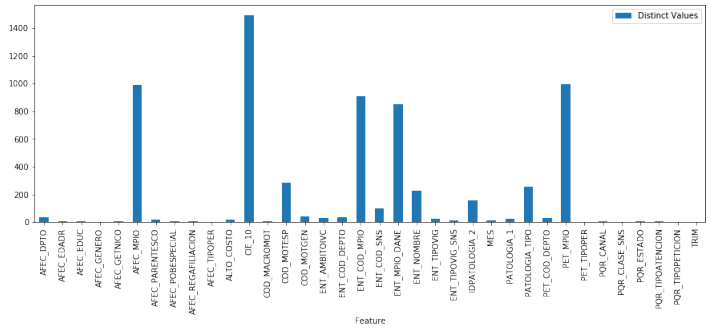
## 7.3 Redundant features

These features represent the same data (100% correlation), so we can keep only the codes and loose the description.

- COD_MACROMOT is the code for MACROMOTIVO
- COD_MOTGEN is the code for MOTIVO_GENERAL
- COD_MOTESP is the code for MOTIVO_ESPECIFICO
- ENT_COD_DEPTO is the code for ENT_DPTO
- ENT_COD_MPIO is the code for ENT_MPIO
- PET_COD_DEPTO is the code for PET_DPTO

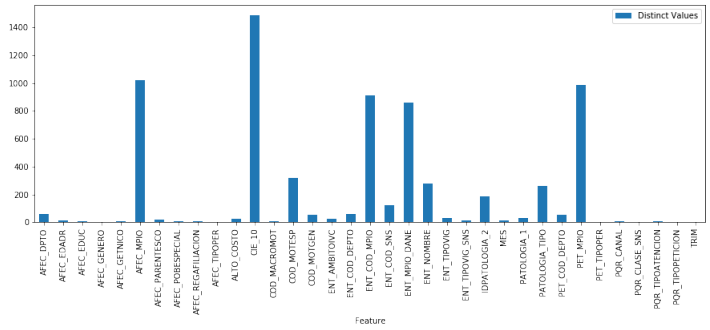
After applying Mean Encode features start to show a more balanced appearance (Figure 5).

Figure 5: Dataset after Mean Encoding




QuantileTransformer slightly improve statistic value for feature AFECT_DEPTO (Figure 6).

Figure 6: Dataset after QuantileTransformer



## 7.4 Post processing analysis

After completing the prepossessing step, more positive correlations appear from the dataset (Figure 7). The target feature RIESGO_VIDA still has low correlation weight in relation with the other features, but nevertheless it has more relevance than the correlation analysis in the original dataset.

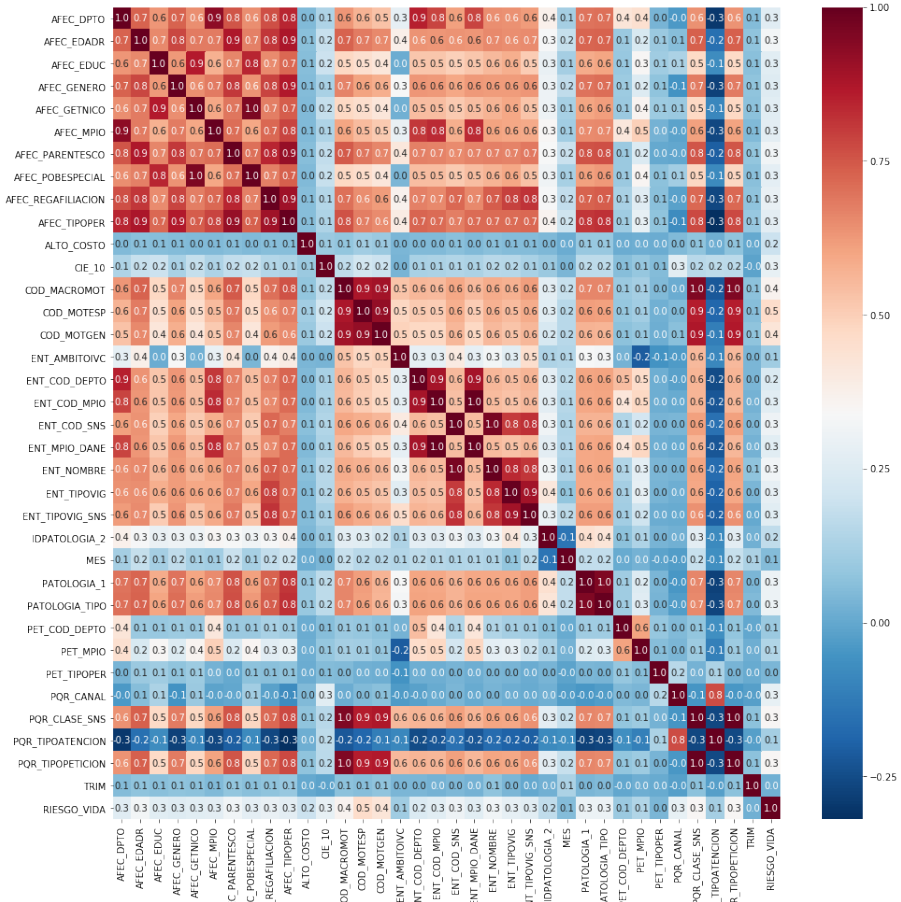
Specifically, the features more related with the target feature are the motive fields (COD_MACROMOT (0.4), COD_MOTGEN (0.4) and COD_MOTESP (0.5)), these are features representing the reason on why the patient is raising the right of request. Some examples of these reasons are:

- Delay in authorization for a surgery
- Deficiencies in patient safety
- Services not adapted to the cultural and particular needs of the population
- Mental health
- HIV and other STDs
- Dehumanized treatment

Although this relationship makes sense, it would be also reasonable to expect the same or even higher correlation with features related with the nature of the illness (CIE_10 (0.3), PATOLOGIA_1 (0.3), IDPATOLOGIA_2 (0.2)), but these features have in fact less correlation value with the target feature RIESGO_VIDA. This might suggest that the reason for which a life's patience may be at risk is more related with administrative issues than the nature of the illness itself.

But it's also worth to point out that motive fields seems to cover not only administrative reasons that may put the patient life at risk but also cultural and illness related causes, so in the end, it seems that these features (COD_MACROMOT, COD_MOTGEN and COD_MOTESP) are trying to cover a considerably wide spectrum of causes for raising a right of petition.

Figure 7: Correlation map



# 8 Implementation

After the dataset processing the target feature still has low correlation with most features, so two ensemble methods where evaluated to boost the model. A third GaussianNB classifier also was evaluated to have a reference at the moment of choosing which learner to use in the final model.

First the dataset was split in training and testing with a training size of 15%. The table show the best scores from best to worst (Table 1).

Table 1: Learners scores

| | Learner | F Beta Score = 2 |
|---|---|---|
| | RandomForestClassifier | 0.727557 |
| | AdaBoostClassifier | 0.657321 |
| | GaussianNB | 0.633024 |

# 9 Refinement

Next, Hyper-parameter optimization was applied for the two best model's scores.

## 9.1 Grid search for RandomForestClassifier

```
rfParameters = {
  'criterion':['gini', 'entropy'],
  'max_depth':[5, 10, 15],
  'max_features':['auto', 'sqrt', 'log2', None],
  'class_weight': ['balanced', 'balanced_subsample'],
}

Unoptimized model
------
F-score on testing data: 0.7276

Optimized Model
------
Final F-score on the testing data: 0.8015
```

## 9.2 Grid search for AdaBoostClassifier

```
adaParameters = {
  'learning_rate':[0.1, 0.5, 1],
  'algorithm' :['SAMME', 'SAMME.R']
}

Unoptimized model
```

```
F-score on testing data: 0.6573

Optimized Model
------
Final F-score on the testing data: 0.6573
```

### 9.3 Stacking with two best classifiers

F-score on StackingClassifier with rfClassifier and adaClassifier: 0.7298

### 9.4 Stacking with all classifiers

F-score on StackingClassifier with rfClassifier, adaClassifier and GaussianNB(): 0.7298

### 9.5 Over sampling over best classifier

Oversampled score: 0.801487

## 10 Model Evaluation and Validation, Justification

After pre-processing the data and using several strategies for improving a model, RandomForestClassifier gave the best score of 0.8007, outperforming the benchmark model that had a score of 0.5553 (Table 2).

Table 2: Models score evaluation

| | Benchmark | Unoptimized Model | Optimized Model |
|---|---|---|---|
| | 0.5553 | 0.727557 | 0.8015 |

Although the the final model managed to reach a score about 80%, it's still a low score given the fact that the goal is about detecting people with life at risk.

## 11 Conclusions

The dataset has very imbalanced data, both in the features and target features. From the 2'130.783 records, about 15% has a target data of 1, that is, patients with life at risk. Another challenge is the low correlation between the target label and the dataset features. The pre-processing step allowed a more balanced behaviour with the dataset features, but the model still struggled to reach a good performance.

Something worth mentioning is the fact that features COD_MACROMOT, COD_MOTGEN and COD_MOTESP had very low correlation with the target

feature RIESGO_VIDA before the cleaning / pre-processing step, moreover, these fields had the lowest correlation values with the target score. Nevertheless, these same features gain more correlation weight over all the features after processing the dataset and became the most important features to make better predictions.

Several strategies were taken into account: Hyperparameter tuning, over-sampling and stacking, being Hyperparameter tuning the one that gave better results with a score about 80%. The model was created using a Intel Core i7 2,9 GHz with 16 GB 2133 MHz LPDDR3 of RAM. More hyper-parameters would considerably slow down the model building, so small subsets of parameters were used for the hyperparameter tuning. A more extensive set of parameters could be used if better hardware available, same for re sampling with other approaches like ClusterCentroids, NearMiss-3 or AllKNN: These algorithms were tested but leaved out due to the large computing time consumed.

# References


[1] Colombia Government. *Datos Abiertos. Govierno Digital*. URL: `https://www.datos.gov.co/en/`.

[2] Colombia Government. *FAQ Derecho de Petición*. URL: `http://wp.presidencia.gov.co/sitios/dapre/atencion/Paginas/preguntas-frecuentes.aspx`. (accessed: 29.11.2018).

[3] Superintendencia Nacional de Salud. *Base De Datos PQRD 2015*. URL: `https://www.datos.gov.co/Salud-y-Protecci-n-Social/Base-De-Datos-PQRD-2015/36n3-fsjh`. (accessed: 29.11.2018).

[4] Superintendencia Nacional de Salud. *Base De Datos PQRD 2016*. URL: `https://www.datos.gov.co/Salud-y-Protecci-n-Social/Base-De-Datos-PQRD-2016/b3xk-8uh2`. (accessed: 29.11.2018).

[5] Superintendencia Nacional de Salud. *Base De Datos PQRD 2017*. URL: `https://www.datos.gov.co/es/Salud-y-Protecci-n-Social/Base-De-Datos-PQRD-2017/gg2r-kx6x`. (accessed: 29.11.2018).

[6] Superintendencia Nacional de Salud. *Supersalud Contact website*. URL: `https://www.supersalud.gov.co/es-co/atencion-ciudadano/contactenos`.