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

Mario Rugeles Pérez

December 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) [7].

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 ([4], [5], [6]) 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.

This is the strategy for the pipeline's construction:

2.1 Data Analysis

A first look to the data will show the characteristics of the dataset that will serve to decide how to prepocess the data

2.2 Preprocessing

The dataset will be prepossessed to unify the features: the dataset is a union of separate datasets per year (one for 2015, another for 2016 and the last for 2017). Every year a few features can be added or removed, so the first step will be identify all the features these three datasets have in common and create a new dataset with these selected features.

2.3 Building the model

After the preprocessing step the model will be build with three different classificators. The selected classificators will depend of the observations made on the dataset.

2.4 Tuning

Hyperparameter tuning will be applied on the selected classificators.

2.5 Validation

The best classificator's score will be compared against the Benchmark's score and also tested against a validation set that is not part of the trainig nor the test set.

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.

F-beta score allows to balance precision against recall in just one formula, lower beta (0.5) will favor precision, while higher beta (2) will give more weight to recall.

$$F_\beta = (1 + \beta^2) \times \frac{precision \times recall}{(\beta^2 \times precision) + recall}$$

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.

Figure 1: Sample from the dataset

	AFEC_DPTO	AFEC_EDADR	AFEC_EDUC	AFEC_GENERO	AFEC_GETNICO	AFEC_MPIO	AFEC_PARENTESCO	AFEC_POBESPECIAL	AFEC_REGAFILIACIO
0	BOGOTA D.C.	DE 13 A 17 AÑOS	Ninguno	Mujer	No aplica	BOGOTÁ	Otro	No aplica	Subsidia
1	NARIÑO	DE 0 A 5 AÑOS	Ninguno	Hombre	No aplica	PUERRES	Otro	No aplica	Subsidia
2	VALLE	DE 13 A 17 AÑOS	Secundaria	Hombre	No aplica	CALI	Abuelo (a)	No aplica	Contributi
3	HUILA	DE 13 A 17 AÑOS	Universitario Incompleto	Hombre	No aplica	NEIVA	Padre	No aplica	Contributi
4	RISARALDA	DE 13 A 17 AÑOS	Ninguno	Hombre	No aplica	DOSQUEBRADAS	Padre	Persona en Condición de Discapacidad	Contributi

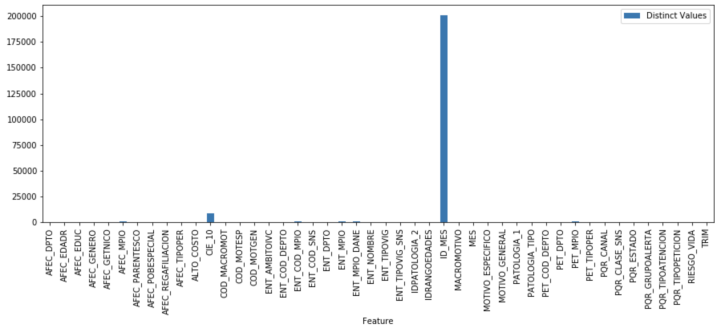
5 rows × 46 columns

The dataset has 45 features plus the target feature with 2'375.371 records that will split in training, test and validation sets.

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.

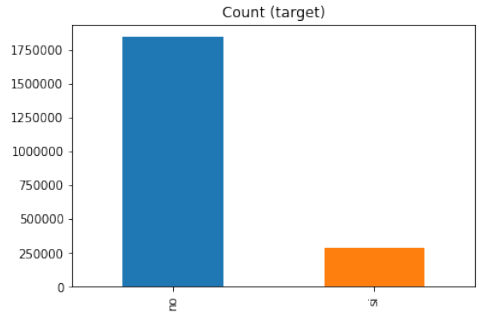
A first visualization shows high imbalanced categorical features with high cardinality (Figure 2).

Figure 2: High cardinality



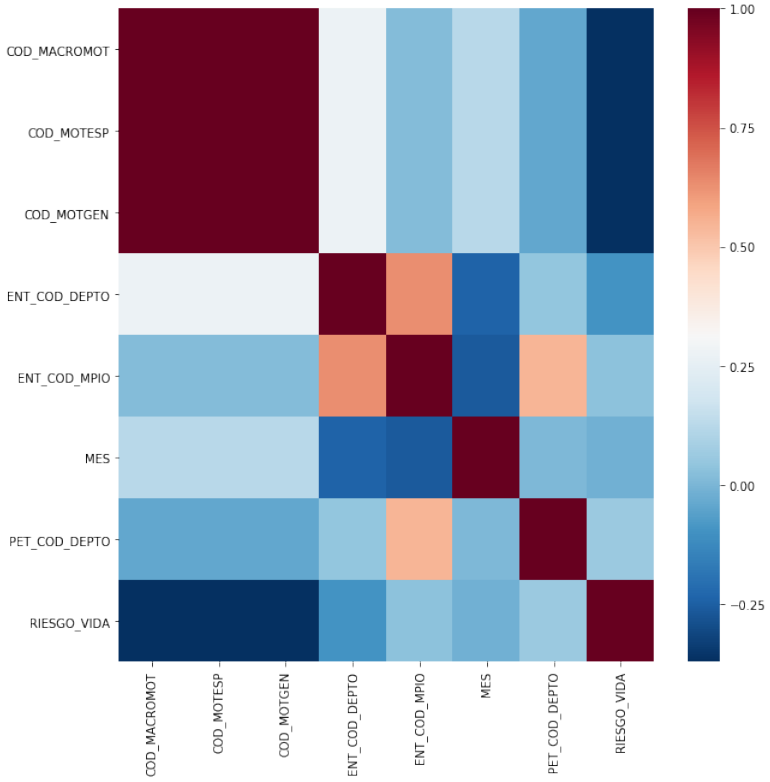
The labels also show a high uneven distribution (Figure 3). Re-sampling will be taken into account in the dataset prepossessing.

Figure 3: 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 4).

Figure 4: Low correlations with target feature RIESGO_VIDA



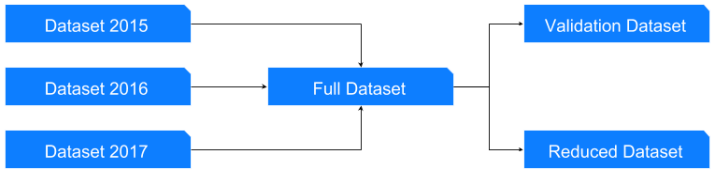
4.1 Missing data

A more detailed observation of the dataset shows several records with missing information that can impact the model performance. For that reason, two datasets were created, a full dataset that ignores the fact of records with missing data, and a reduced dataset with records with no more of 3 missing data per row. The reason of choosing records with no more of 3 missing data per row for the reduced dataset was to avoid a big reduction of records for the dataset that can also impact the model's performance (Figure 5, Figure 6).

Figure 5: Missing Data

	AFEC_DPTO	AFEC_EDADR	AFEC_EDUC	AFEC_GENERO	AFEC_GETNICO	AFEC_MPIO	AFEC_PARENTESCO	AFEC_POBESPECIAL	AFEC_REGAFILI
632038	bogota d.c.	de 30 a 37 años	ninguno	mujer	no aplica	bogotá	nombre propio	no aplica	contr
10533	valle del cauca	de 50 a 62 años	0	mujer	0	calcedonia	nombre propio	0	régimen e
21130	0	de 6 a 12 años	0	mujer	0	0	padre	0	contr
295862	santander	de 25 a 29 años	0	hombre	0	bucaramanga	conyugue	0	contr
359762	0	0	0	0	0	0	0	0	
604173	0	0	0	0	0	0	0	0	
599009	0	0	0	0	0	0	0	0	

Figure 6: Dataset build pipeline



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 was not be used in this project as it would 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, so a clean process was made to correct this case.

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

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

Given the low correlation with the target variable, and the fact that there are considerably many features, the model required 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.

To help to increment the score, the ensemble learners RandomForestClassifier and AdaBoostClassifier were selected due to the relatively low speed for training with both default parameters and grid search.

RandomForestClassifier helps to minimize the over fitting issue with Decision Trees and works well with large datasets (The dataset for this project is greater than one million records).

AdaBoostClassifier uses few parameters so it can be easier to configure and is designed to boost decision trees performance on binary classifications which makes it a good fit for this project.

Gaussian Naive Bayes was considered because it deals with real values and the dataset is composed of real values after the preprocessing.

The data was split in training, testing and validation datasets. Process of training and tuning was made over the training and testing dataset and then the models where exposed to unseen data with the validation dataset. This process was performed for both the full dataset and the reduced dataset.

6 Benchmark

A Logistic Regression model will be use as a base model to compare with the final model because it demonstrates the need to implement a non linear solution to the problem.

7 Data Preprocessing

7.1 Removing fields

According to the official 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).

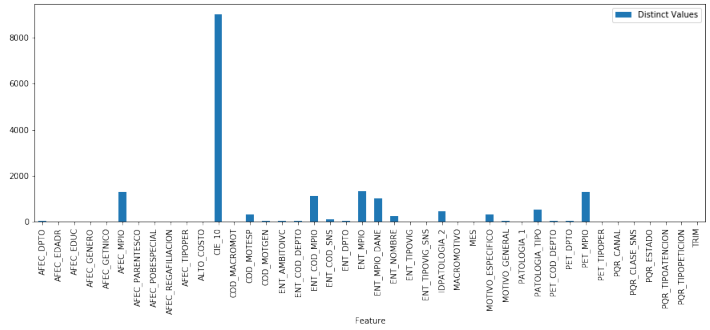
7.2 Special cases: AFEC_DPTO, PET_DEPTO

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 relatively 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 7).

Figure 7: Some features gain statistical meaning



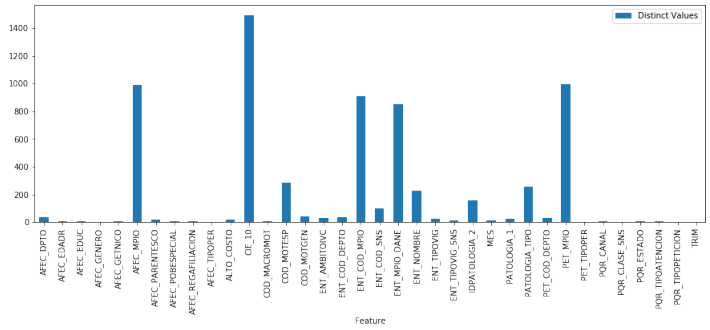
7.3 Redundant features

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

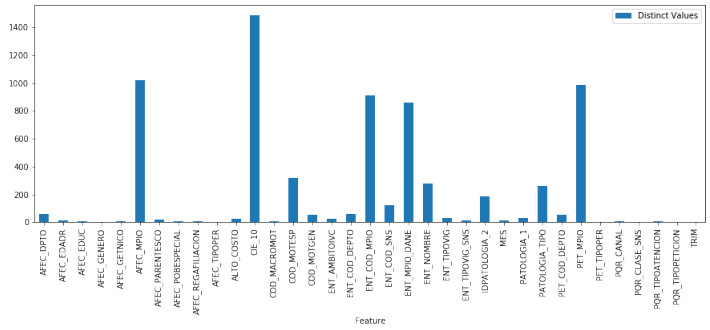
After applying Mean Encode features start to show a more balanced appearance (Figure 8).

Figure 8: Dataset after Mean Encoding



QuantileTransformer slightly improve statistic value for feature AFECT_DEPTO (Figure 9).

Figure 9: 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.

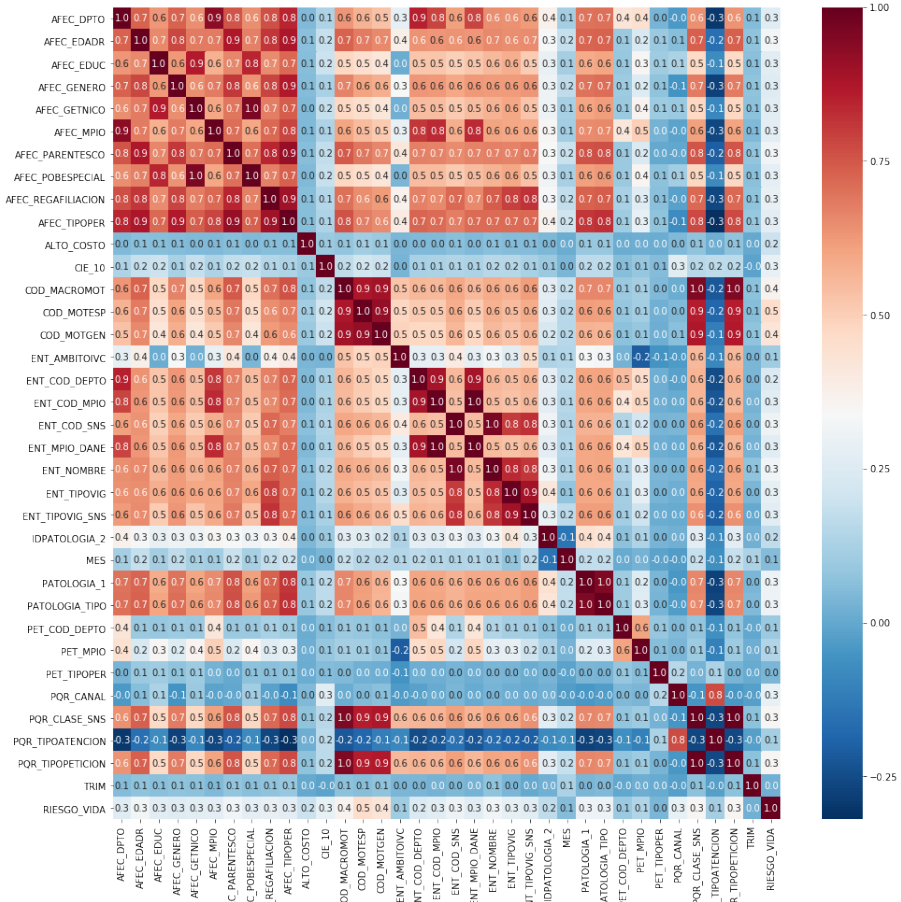
Specifically, the features more related with the target feature are the motive fields (COD_MACROMOT (0.4), COD_MOTGEN (0.4) and COD_MOTESP (0.5)), these are features representing the reason on why the patient is raising the right of request. Some examples of these reasons are:

- 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 10: Correlation map

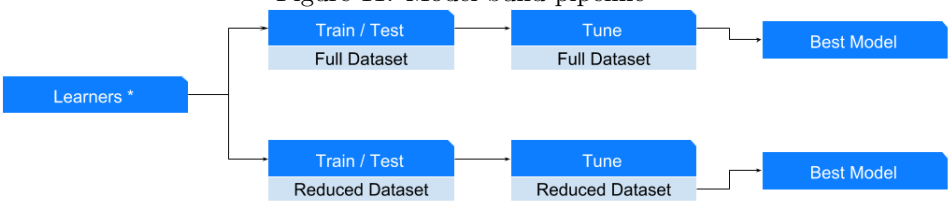


8 Implementation

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 because although is not a ensemble methods it performs well with real value data and could give a boost in the stacking method.

Two pipelines were implemented, one for each dataset. The first dataset keeps records with rows with four or more features with value equals to 0, and the second dataset with rows with no more of three values equals to zero.

Figure 11: Model build pipeline



The table show the best scores from best to worst with the full dataset (Table 1).

Table 1: Learners scores

Learner	Full Dataset Score	Reduced Dataset Score
RandomForestClassifier	0.760025	0.716901
AdaBoostClassifier	0.703412	0.714472
GaussianNB	0.524448	0.664567

All models performed with quite low scores with default parameters. The models trained with the reduced dataset shown slight low values on average but in general all boost models were around 70% - 75% score.

9 Refinement

Next, Grid search, Stacking and Oversampling optimization was applied for the two best model's scores. Hyper-parameter optimization was also tested, but it took too long to complete with the hardware available and was discarded for this project (Table 2).

After running several strategies for improving a model, RandomForestClassifier gave the best score of 0.7986 with the reduced dataset, outperforming the benchmark model that had a score of 0.5770 (Table 3).

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.

Table 2: Tuned models

Learner(s)	Method	Full Dataset Score	Reduced Dataset Score
RFC	GridSearchCV	0.7956	0.7986
RFC	RandomOverSampler	0.7961	0.7264
RFC, Ada	Stacking	0.7220	0.7242
RFC, Ada, Gauss	Stacking	0.7205	0.7264
Ada	GridSearchCV	0.6568	0.6646

Table 3: Models score evaluation

	Benchmark	Reduced Dataset Score
	0.5770	0.7986

10 Detour from the original plan: Exploring a different approach

The initial idea was to choose some learners based in the properties of the processed dataset, run several tuning approaches, select the best model and perform some validation to the best model. Nevertheless, although the best model outperformed the benchmark by about 38% it still had a relatively low score for the task at hand.

Other options were to use another learners suitable for non linear problems, like SVM, but these approaches turned out to consume too much time and resources. So a new approach was considered.

10.1 Considering poor dataset's information

As the dataset has notably low correlations between features an target, that posed a challenge to the tested algorithms to make good predictions. Before concluding the possibility that the dataset may had not enough information to allow good predictions, a new hypothesis was considered: Maybe it can be more information that can not be deduced by 'classical' algorithms or takes a lot of resources for doing that.

10.2 Dataset as two dimensional images

CNN's are able to build an internal representation of 2 dimensional images [3]. That means the features can be evaluated according their value and position in the array, allowing the algorithm to find new correlations between the features and the target value.

Transforming the features as 2 dimensional images (with no rgb dimensions) reveals a new pattern: Records representing patients with life at risk show predominantly higher overall values, which is represented as lighter average color per image (Figure 12, Figure 13).

Figure 12: 20 Records representing patients with no life at risk

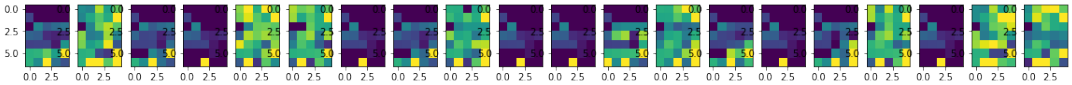
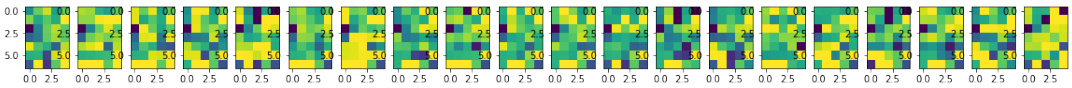


Figure 13: 20 Records representing patients with life at risk



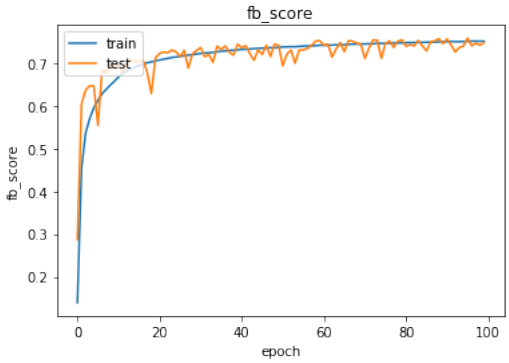
The final model uses 2 Convolution layers with Max Pool Layers each one with a fb-score of 0.9391 (Listing 1, Figure 14).

Listing 1: Final Model

Layer (type)	Output Shape	Param #
conv2d_22 (Conv2D)	(None, 7, 5, 16)	80
conv2d_23 (Conv2D)	(None, 7, 5, 16)	1040
max_pooling2d_8 (MaxPooling2	(None, 3, 2, 16)	0
conv2d_24 (Conv2D)	(None, 3, 2, 32)	2080
conv2d_25 (Conv2D)	(None, 3, 2, 32)	4128
max_pooling2d_9 (MaxPooling2	(None, 1, 1, 32)	0
global_average_pooling2d_11	(None, 32)	0
dense_16 (Dense)	(None, 2)	66

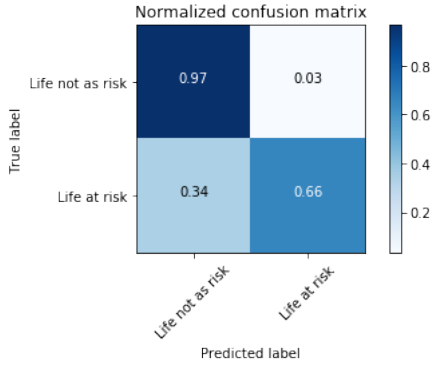
Total params: 7,394
Trainable params: 7,394
Non-trainable params: 0

Figure 14: Learning Curve for CNN



The model had a score of 0.7329 and a high value of false negatives (34%), not suitable as a solution for this problem. So although this approach was abandoned, it was part of the whole project for finding a good prediction model and it was finally included in this report.

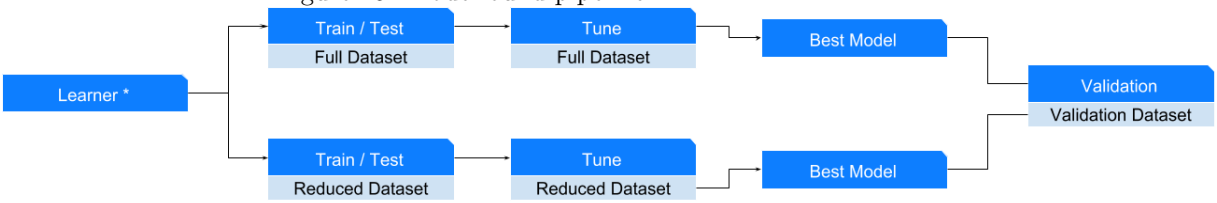
Figure 15: Confusion Matrix with Validation Data



11 Model Evaluation and Validation, Justification

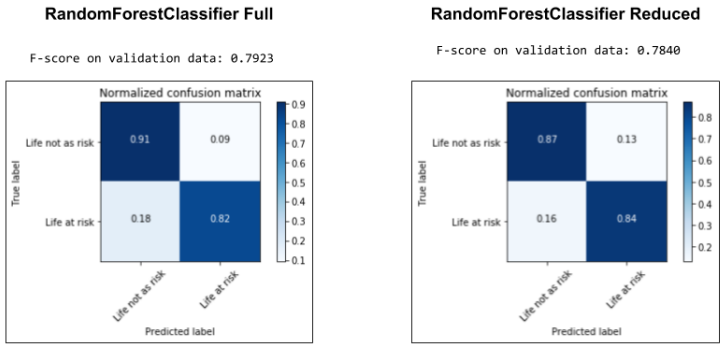
The final part of the Model Building Pipeline was to test all models with unseen data (Validation Dataset) (Figure 16).

Figure 16: Model build pipeline



Again RandomForestClassifier's models gave the best results, in both models (The one created with the full dataset and the second created with the reduced dataset)

Figure 17: Model Validation



The model trained with RandomForestClassifier with the full dataset had a better score of .7923 with a false negatives percentage of 18%. Model trained with the reduced dataset had a slight lower percentage of false positives of 16% but also had a notably higher value of 13% of false positives, compared with the 9% of the first model. That will increase the cases of false alarms that the users of the model will have to respond, that make the model created with the reduced dataset not viable because although we want less mistakes leaving patients with life at risk undetected we also to reduce work load over the government agents by reducing the probabilities of giving priority to false positives.

12 Conclusions

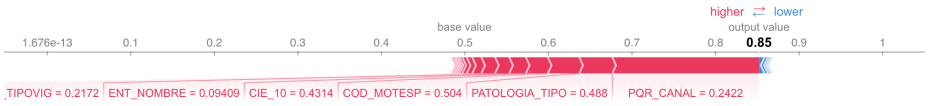
The dataset has very imbalanced data, both in the features and target features. From the 2'130.783 records, about 15% has a target data of 1, that is, patients with life at risk. Another challenge is the low correlation between the target

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.

A closer look to the force plot shows the relevance of features COD_MOTESP (Reason for the right of request), PATOLOGIA_TIPO (Pathology type) and PQR_CANAL (Communication channel) which makes sense as the decision of whether or not a patients life is at risk of not should be related with the patient's pathology. (Figure 18)

Figure 18: Force Plot



It's also worth pointing out that almost every feature contributes to the model's prediction.

Overall, all models gave an score around 70% (Including the CNN model) and 80% (For the tuned ones) which might suggest all models were converging to similar solutions, so it's worth considering the this might be the best the data can offer. The real problem is why the low correlation mentioned earlier. A letter was send to the Supersalud asking on how the field RIESGO_VIDA (Life at risk) is determined and as of the time writing of this report no response has been send. Further analysis should be done when new information from the Supersalud is available.

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] PhD Jason Brownlee. *When to Use MLP, CNN, and RNN Neural Networks*. URL: https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/. (accessed: 05.12.2018).

[4] 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).

[5] 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).

[6] 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).

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