**Asthma Recurrence and Associated Risk for Omnivida S.A.S**

**Final Report Project.**

**Colombia | Team #59**

**Andrés García, Andrés Mejía, Ariel García, Catalina Moreno, Gilberto A. Morales Zamora, Sergio Gómez**

# Summary

Adherence is critical for asthma patients, as their condition can worsen quickly if they stop taking their medication. To help Omnivida to improve the quality of life of their patients, we aim to provide this company insights on the principal aspects that have incidence in the patient’s adherence. This document presents the analysis of variables obtained by crossing the different bases of the study, to find aspects or characteristics of the patient that explain the adherence to the medication.

# Business Problem

Omnivida is a health risk management company that integrates patient monitoring, different health system actors, and technology to add value to the health system. Amongst patients that require monitoring, those with chronic asthma are of special interest. The outcome for these patients depends not only on the health system and practitioners’ quality but also on the patient’s willingness to effectively follow instructions related to his condition. The patient’s adherence or compliance is the degree in which the patient correctly follows the medical advice. As long as there is a proper estimation of the adherence of a patient, before the start of the treatment, every agent in the system (health personnel, health provider institutions, and health insurers) can refine best practice policies that help the risks associated to non-adherent patients.

The number of attended and detected asthma patients has increased during the last four years, according to the official records of the Colombian Ministry of Health and Social Protection (MSPS) (Figure 1). This can be considered as an approximation of the prevalence of Asthma in Colombia, nevertheless, the increase of detected cases could also reflect an improvement of the coverage given by the health system. These databases have the bias of recollection, as they only record those patients that access the public system and not the services given by doctors in private practice.

The increase of the coverage in the health system is also reflected in the reduction of asthma-related deaths recorded by the National Administrative Statistics Department (DANE) with a reduction of more than 50% in the last year compared to the records of 2007. In this case, the bias of recollection is reduced as the DANE recollects information not only from the MSPS records but also from several other sources around the country.

|  |  |
| --- | --- |
| Figure I: Number of asthma patients attended per Year  Source: Registros Individuales de Prestaciones en Salud [[1]](#footnote-0) | Figure II: Number of asthma-related deaths per year    Source: DANE Estadísticas Vitales (Defunciones)[[2]](#footnote-1) |

While studying the number of cases in the year 2019, we observe that most cases are concentrated in Antioquia followed by the city of Bogotá D.C., and Atlántico. In the case of Antioquia and Bogotá, both of them concentrate the higher amount of population, and therefore a higher amount of cases is expected (Figure III).

|  |
| --- |
| Figure III: Number of asthma patients attended in 2019 by department    Source: Registros Individuales de Prestaciones en Salud[[3]](#footnote-2) |

In a second approach, that considers the prevalence of each department (Figure IV). Considering a rate per 1000 inhabitants in the region, we find that the departments with the highest prevalence are Atlántico, Bolivar, and in the third-place: Antioquia. Bogotá now occupies the 10th place, which shows that the cases are mainly based on the density population instead of a high percentage.

|  |
| --- |
| Figure IV: Prevalence estimate (number of cases per 1000 habitants) in 2019 by department    Source: Own calculation based in Registros Individuales de Prestaciones en Salud[[4]](#footnote-3) and DANE Proyecciones Poblacionales[[5]](#footnote-4) |

# Patient Characterization

# Demographic Aspects

Asthma has a higher prevalence in the female population. Our base reflects that same ratio with a 2.3:1 ratio of female vs male cases. Also, as we are considering the latest state of a chronic disease, the profiled patients are mostly adults (506 patients) or elderly (158 patients) which jointly represent 76% of the total population considered.

|  |  |
| --- | --- |
| Figure V: Patients according to sex | Figure VI: Patients according to lifecycle |

Given Colombian context and geographical restrictions, people in urban areas tend to have fewer restrictions to access the system. Most of the patients are from the region of Antioquia (80.1%). This last result is expected due to the source of the data: an Antioquia based health provider institution.

|  |  |
| --- | --- |
| Figure VII: Patients according to their living zone | Figure VIII: Patients according to the department of residence |

In terms of weight and height measurements, it is important to mention that these indicators have a regular behavior (159 of height and 64 kg of weight) respect to the Colombian average which wich is 166 of height and 73 kg of weight. However, patients in overweight have the second-largest share (35%), which means it could be highly related to be an asthmatic patient.

|  |  |
| --- | --- |
| Figure IX: Patients according to the body mass index (BMI) classification | Figure X: Patients weight to height relation |

In terms of the patient’s pathological history, the two main groups of diagnostics are related to the respiratory system and skin and subcutaneous tissue. Both groups are consistent with allergic reactions: dermatitis and rhinitis, also considering that asthma is reflected as an allergic reaction in most patients.

|  |
| --- |
| Figure XI: Patients’ comorbidities based on body systems |

The ACT (*Asthma Control Test*) is defined as a control test for asthmatic patients for 4 months. In this test, it is evaluated, by a general set of questions, how the patient feels for its asthmatic conditions, and how these have (or not) affected his/her life.

This data set shows a big picture of a patient's evolution with time, there’s a variable which permits to find how a patient has controlled (or not) asthma over time. This variable (called *test result)* is classified by three different levels: 1) not controlled (coded as 0-red), 2) partially controlled (coded as 1-blue) and 3) fully controlled (coded as 2-green).

|  |
| --- |
| Figure XII: ACT counting made year-by-year. |

According to Figure XII, there are interesting observations:

* The ACT result may vary over time for a single patient, this makes sense as asthma is a chronic condition (doesn’t cure by a specific time-moment)
* The proportion of patients which results in no-control and partially-control seems to be the same over time
* The proportion of patients who have full control over the asthma symptoms are really low over time. This reinforces the idea of making better health campaigns to improve adherence to different treatments

# Adherence Definition

## Self Reporting Adherence

The company carries out a follow-up to its patients through the performance of tests, to corroborate the adherence to medications. There are seven tests carried out: Morisky Green, Smaq1, Smaq2, Espa, nm Espa, Qualitative weighted, and Weighted quantitative. Except for Weighted quantitative, Smap2, and NM Espa, which has a scale of 0 to 6, the rest of the tests have ratings of not applicable, adherence or not.

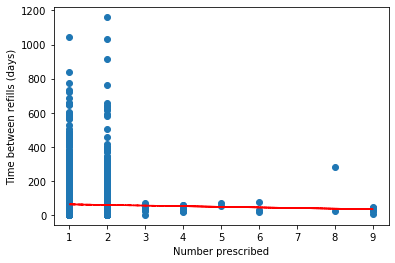
Figure XIII shows the distribution over the entire time horizon of the behavior of the test results. It is possible to verify that in all cases, most of the results show adherence on the part of the patients. It is noteworthy that the results come from surveys, so a verifiable method should be sought to categorize patients into adherents or not.

|  |
| --- |
| Figure XIII: Adherent patients according to self-reported tests |

## Pharmacological tracing

Another approach to this problem is using pharmaceutical records to track the patient’s adherence. This is done by contracting a patient's medication and dosage in contrast to the time between refills. In theory, time between refills and dosage should be highly correlated in adherent patients and not so much in nonadherent ones. We start checking if this approach is plausible by plotting the time between refills and dosage.

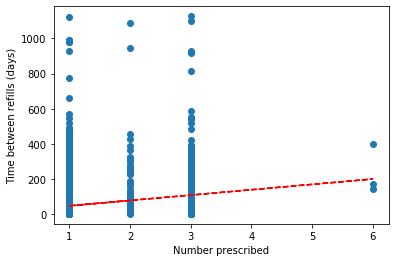
* Salbutamol

Figure XIV: Prescription vs refills plot for Salbutamol.

Salbutamol is mostly used as a crisis inhaler to mitigate asthma attacks, it is commonly formulated as a monthly refill, yet we see that most patients that refilled took more than 200 days between refills, also given its nature (emergencies only not daily use) a refill may not be necessary.

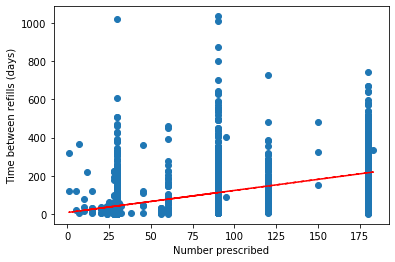
* Fluticasone propionate

Figure XV: Prescription vs refills plot for Fluticasone.



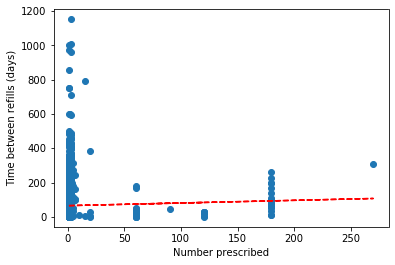
This is a steroid medication used for long term asthma management, this more closely resembles what we would expect from an adherent group of patients, the trend line has a slope of 30.5 and that matches our expectations that an inhaler should last a month.

* Montelukast

Figure XVI: Prescription vs refills plot for Montelukast.

Montelukast is an oral medication used to control asthma attacks, it is used as a daily pill, this medication seems to be a good indicator for adherence, for example, the trend has a slope of 1.15 consistent with daily intake.

* Budesonide

Figure XVII: Prescription vs refills plot for Budesonide.

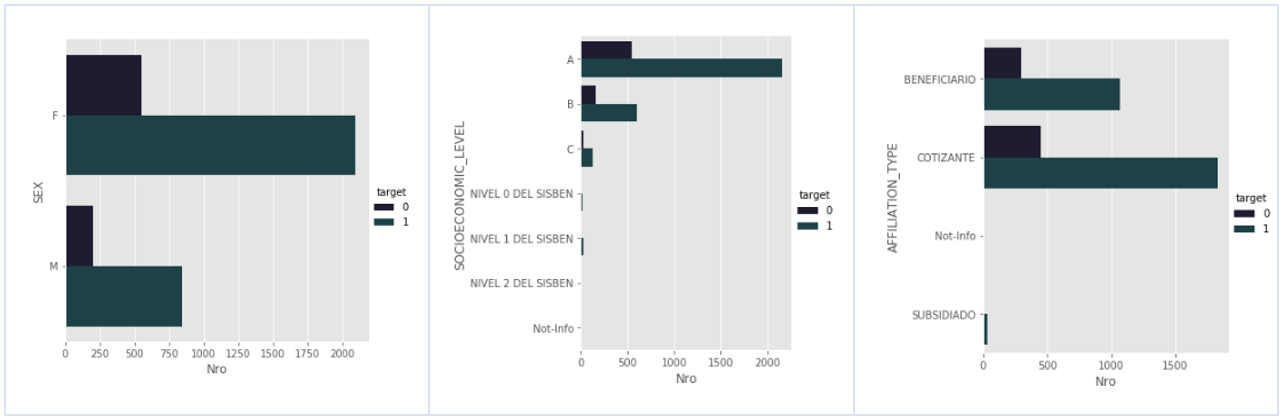
Budesonide is a corticosteroid mostly used as an inhaler for the treatment of asthma, although it is used as a maintenance and prophylactic treatment, it doesn’t seem to follow a close pattern between doses prescribed and time between refills.

Although the use of the time between medications seems promising, there are several confounding variables that we have to take into account, asthma and allergic rhinitis are related in an epidemiologic level, also there is an important overlap in medication, so sometimes the same medication is prescribed for both conditions. This is also an issue as sometimes specialists don’t prescribe medication for one of these conditions if that a homologous one was prescribed by another professional, so in these cases, adherence has to be tracked across conditions.

# Exploratory Data Analysis

The next analyses will concentrate on looking for relationships of the different variables concerning the level of adherence (yes or no) that a patient may present.

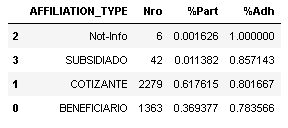
Figure XVIII: Relationships between variables.



Adherence to the medication seems to be uniform throughout a patient's sex. There is no clear pattern identifying which age group is more adherent. It seems that the adherence rate is slightly higher for females than for males with around ~1.0%.



In the health system, there are two types of patients according to the contribution scheme, contributory patients are those who pay for health services monthly, while the beneficiary patient corresponds to a patient affiliated to the health with link blood or affinity with someone who is a contributor.

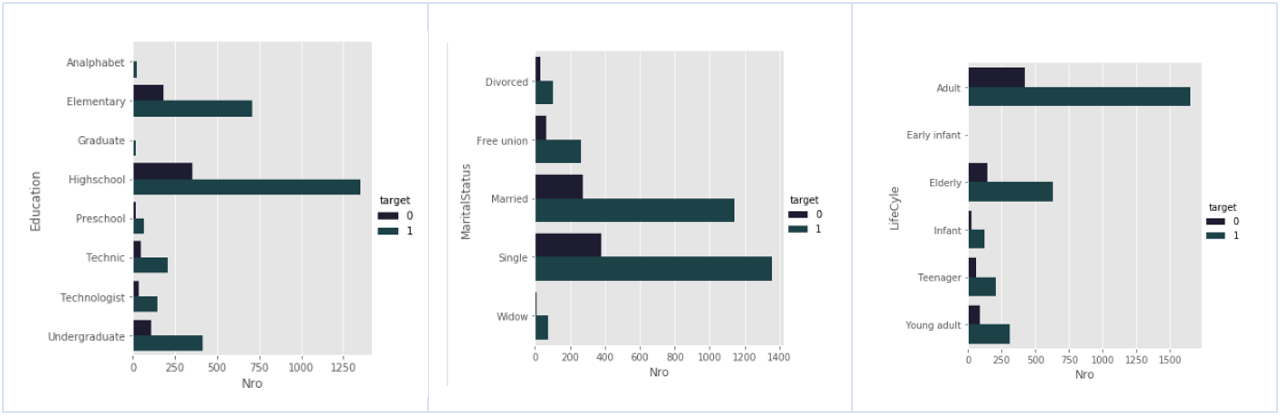


Here it is observed that the majority of patients with chronic asthma are on a contributory (cotizante) regime (61%) and the rest are beneficiaries. We also see a difference in adherence rates of 2-5% between these segments.

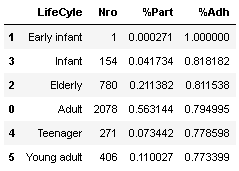
This socioeconomic classification corresponds to a variable generated by the client's company based on income levels. Looking at the level of adherence according to socioeconomic level, in aggregate, a relatively stable adherence close to ~80% can be observed. However, this is a fairly high level of aggregation because these levels include income groups that can be quite large. Also, we can notice that some values have to be grouped to obtain a category with more participation (join all SISBEN values in one). For C socioeconomic level, there is a 1% more adherence rate compared with A and B.



Figure XIX: Relationships between variables an adherence (part 1).



Adherence to the medication seems to be uniform throughout a patient's age. There is no clear pattern identifying which age group is more adherent.



According to the educational level, it can be observed that the levels of adherence for educational levels considered low (preschool, high school, and elementary) are good with ~80% approximately.

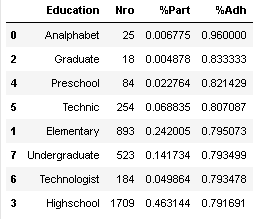
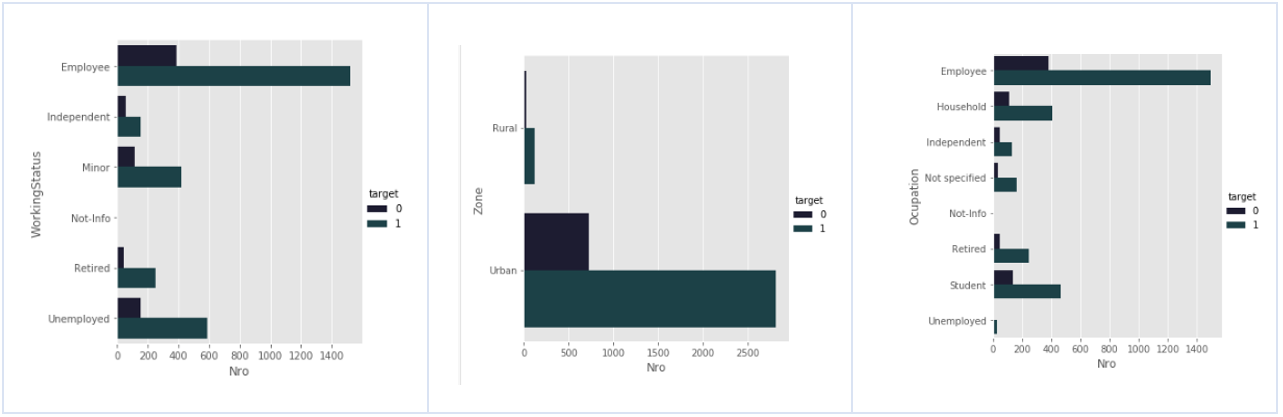
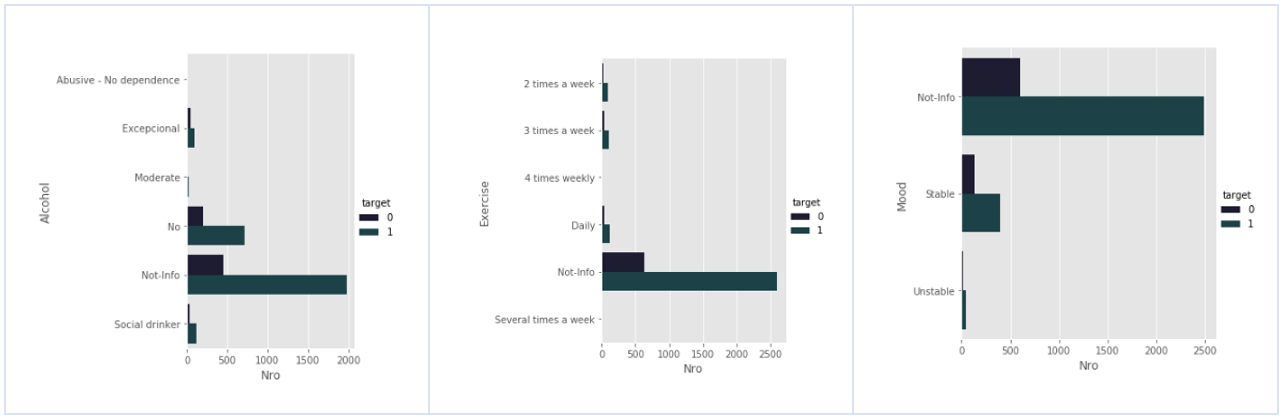
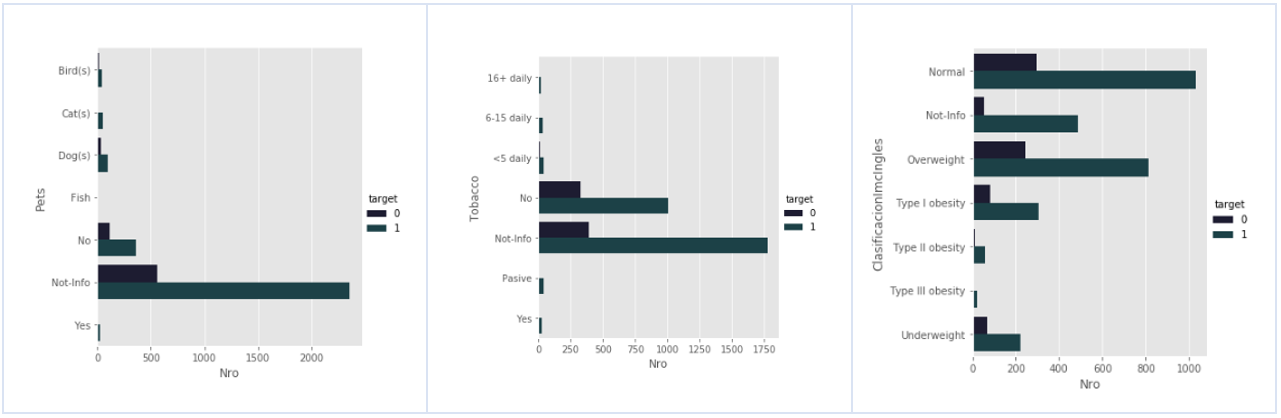


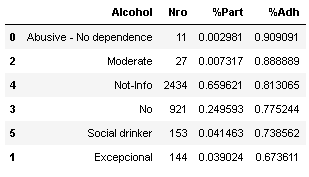
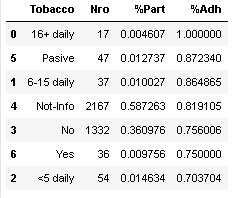
Figure XX: Relationships between variables an adherence (part 2).



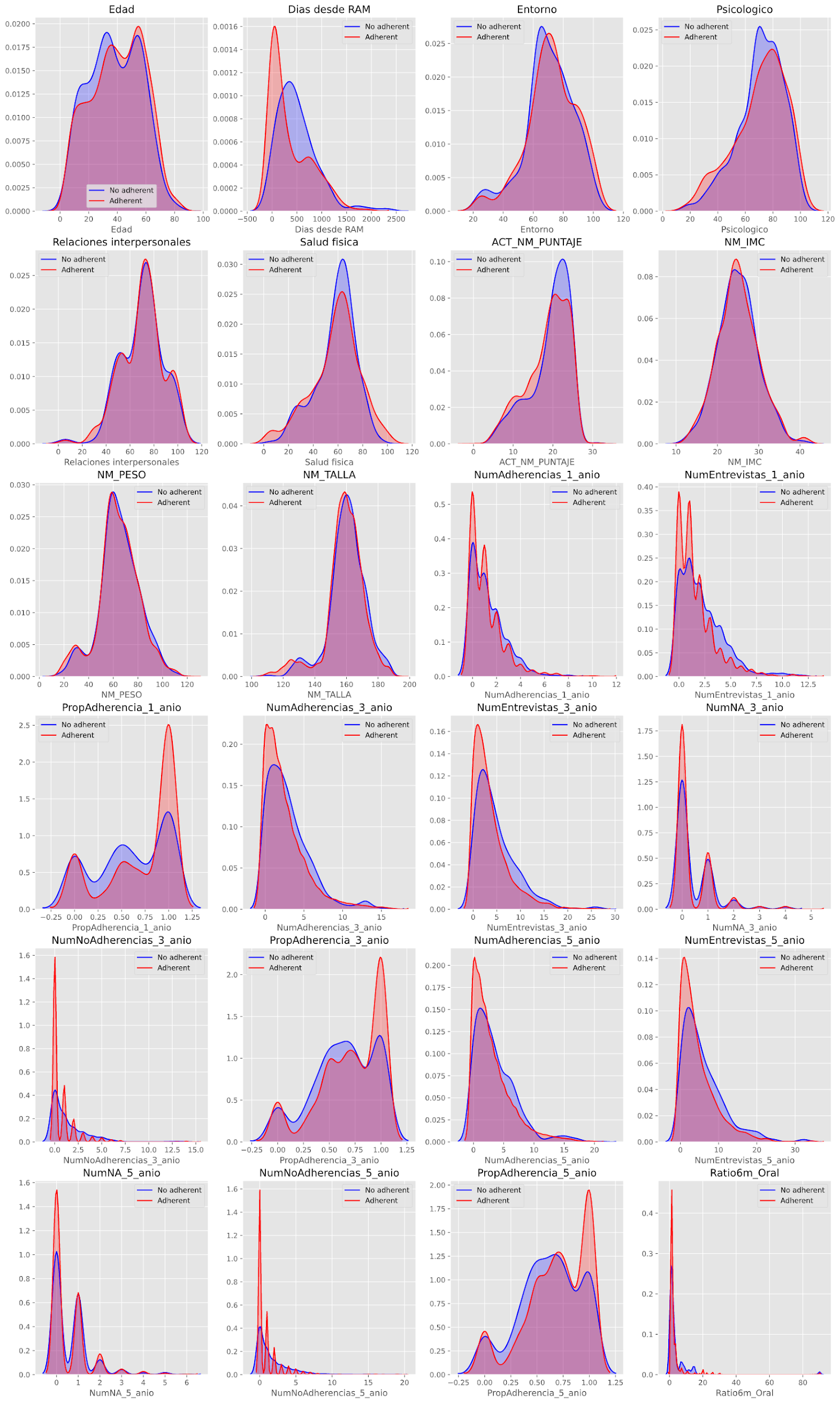


Within the follow-up of patients, we find the habit of consumption, such as tobacco and alcohol. we can see that, although there are few patients with the record, we see a difference between the rates of medication adherence.

Figure XXI: Relationships between variables an adherence (part 3).

In the case of the continuous variables, there is no visible trait that presents an individual significant interaction with the adherence. Nevertheless, we can observe that the population is properly distributed around all population characteristics.

Figure XXII: Population distribution by characteristics.

# Predictive Models

The expected prediction consists of a binary variable of adherence/non-adherence. Taking that into account, the model of choice will be a classification method to predict the probability of adherence to a medication.

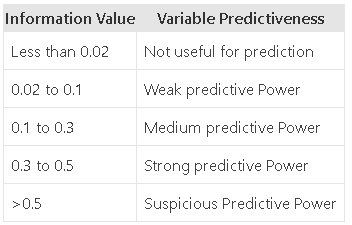
Amongst the possible methodologies to predict the adherence, the ones that appear to be the most appropriate, and therefore, the ones that are going to be considered are:

* Logistic regression model.
* Decision tree-based models:
  + Random Forest.
  + Gradient Boosting.

A first filter was applied to the database to eliminate those variables that not represent relevant informatión, like dates and columns referred to identifications. Before the estimation of the mentioned methodologies, a selection of variables was performed to filter (second filter) non-informative covariates. This was based on the following criteria:

* Percentage (%) of null values.
* Information value (IV)

In the annexed **A,** a list of all the variables included to estimate a model is shown, with their respective missing percentage and the information value, also show is the variable was included in the first and second filter.



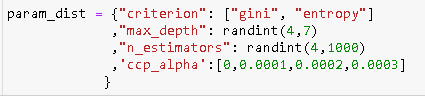
Based on the previous table, a cut-off of 4% for IV was defined to discard not useful covariate for the model. After selecting the covariates taking in count the missing and the IV, a set of 20 features were selected to continue with the model estimation.

In the case of the missing values, validation was performed, to determine if the value imputation is feasible. There were two criteria to impute null values: zero and the median of the variable. Each variable was imputed based on the nature of the variable. With the 20 covariates obtained from the filter one an two, some models were estimated to obtain relevant results.

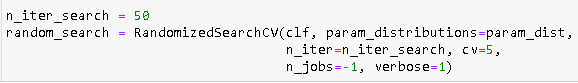
Model Results.

A tree-based algorithm, specifically a random forest was the selected strategy to predict the probability of adherence of a patient; compared to other methodologies, this one presented a better performance in the train and test samples.

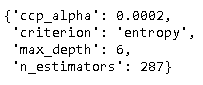
For the estimation of the model, based on the results of the initial selection of characteristics (IV) and the adjustments of the null values ​​and the treatment for the categorical variables, the optimization of the most important parameters of the algorithm was performed using a random search CV.



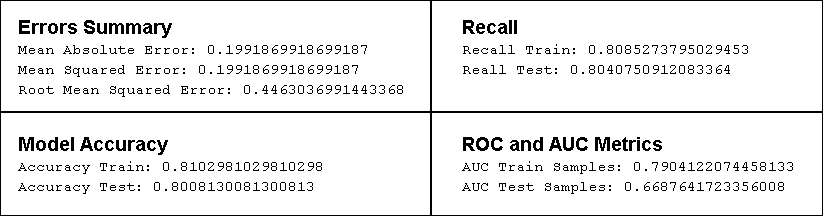
In this exercise, the base was divided into a training sample with 80% of the records for the estimation and optimization of parameters, and 5 k-folds were defined for the cross-validation process.



The optimization results yielded the following parameter configuration:



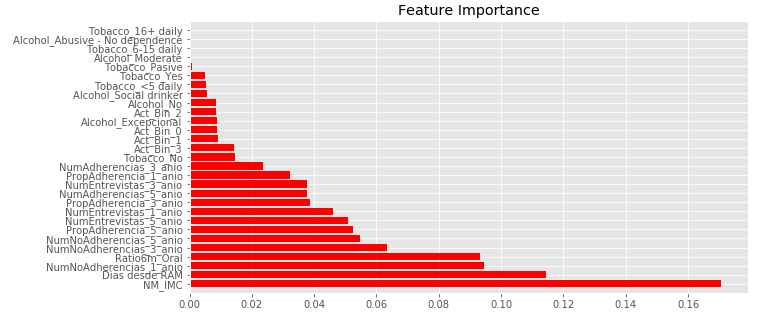
Based on the optimization results, the final model is estimated and the associated estimation metrics are calculated.



**Feature Importances.**

Algorithm based on 287 trees, indicates that BMI (IMC in Spanish) and Days between last negative reaction to the medicament are the most representative covariates to this model, also covariates related to previous adherence interviews (last one, 3 or 5 years) are representative to the model.

Figure XXIII: Characteristic by importance.



Persistence Probability Score.

Based on the previous results of the adherence estimation, we can notice that the model is a good approach of the adherence probability, however, since the measurement of the adherence is based on a weighted average of auto report interviews, we decided to complement the study by providing a model for the claim persistence. The methodology applied for this model was the same as the adherence model. We made a feature selection based on IV, a missing values imputation, and a parameter optimization using RandomizedSearchCV.

Our goal is to provide both scores and build strategies with the medical care experts to make a better management of the health risk. Here are some of the most important results of this model. For this one, a set of 54 covariates was selected to make an Information Value analysis. After this feature selection, 16 covariates were used to perform a Random forest.

**Estimation Metric & Feature Importances.**

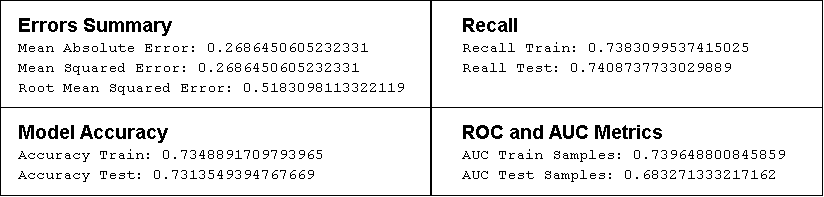
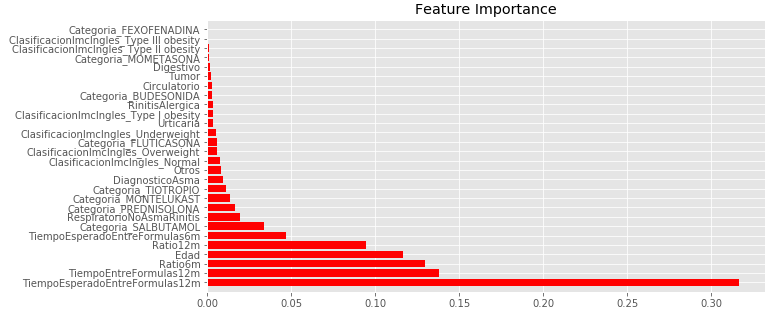
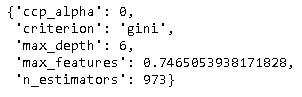
****

Figure XXIV: Estimation metrics by importance.

****

The configuration obtained from the optimization process for this model shows similar parameters compared with the adherence model.



Here, the percentage of features used in each tree is also optimized; for this model the number of trees used for the algorithm is quite large, almost 1000.

# RDS Implementation

An RDS instance was created on amazon AWS, this instance currently hosts the information used to create the dashboard on EC2. Also, two python scripts were written, one for loading new tables into this instance and one with the scripts used by the predictive model and the dashboard to access this data.

We don’t expect this component in the future to need additional setup other than to load additional information needed by either the dashboard or the model.

# EC2 dashboard configuration updates

Three main procedures were followed to dashboard configuration: 1) local configuration of functions, graphics and maps through the integration of Plotly and Dash for Python, 2) local integration of functions and Dash modules, and 3) cloud configuration through Amazon Web Services EC2.

Regarding the first local configuration process using Dash and Plotly, different files were built:

1. index.py which becomes the main file where all sub-functions are called and configured within the main dashboard page.

There is the landing page in which the two modules implemented are related: 1) for health professionals and 2) for health entities. On this page, you will find a brief description of the project and each of the modules mentioned.

Figure XXV: Dashboard landing page.



1. modulo\_1.py is that module where functions are being stored concerning patient information. In this module, you can make a query with the identification number of a particular patient.

Figure XXV: *Module 1 (part 1)*.



This module is intended for health professionals to make a detailed consultation of patient demographic information, evolution in drug claims concerning time, and the calculated probability of a patient being adherent to an asthma treatment.

In general terms, you can find basic patient information such as age, gender, and conditions that may accentuate a patient's non-adherence to an asthma treatment.

Also, you can find the evolution of drug claims for the individual patient concerning different treatments over time. Finally, there is a tachometer-type graph of the likelihood that the individual patient will adhere to an asthma treatment.

Figure XXVI: *Module 1 (part 2)*.

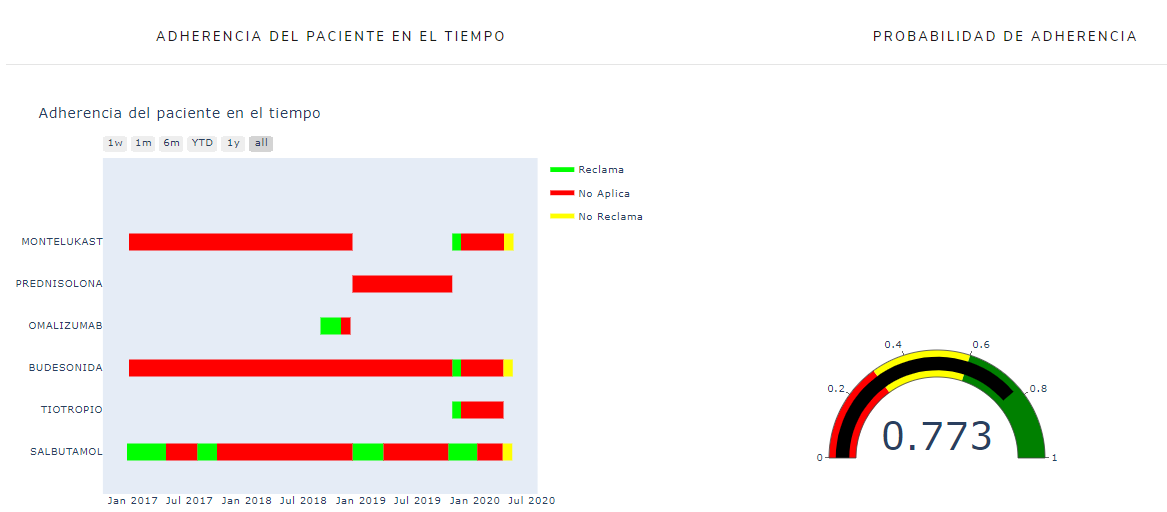
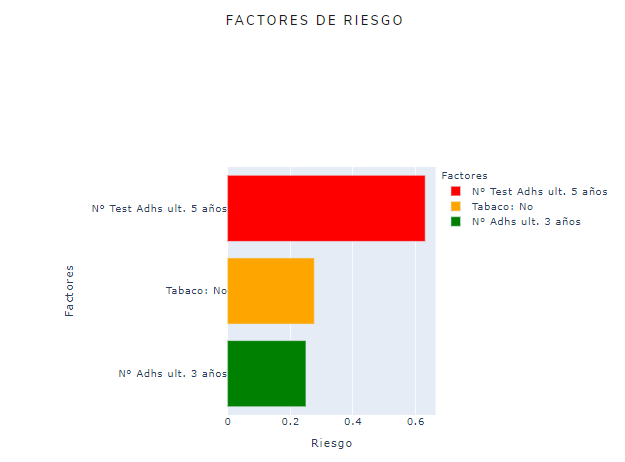


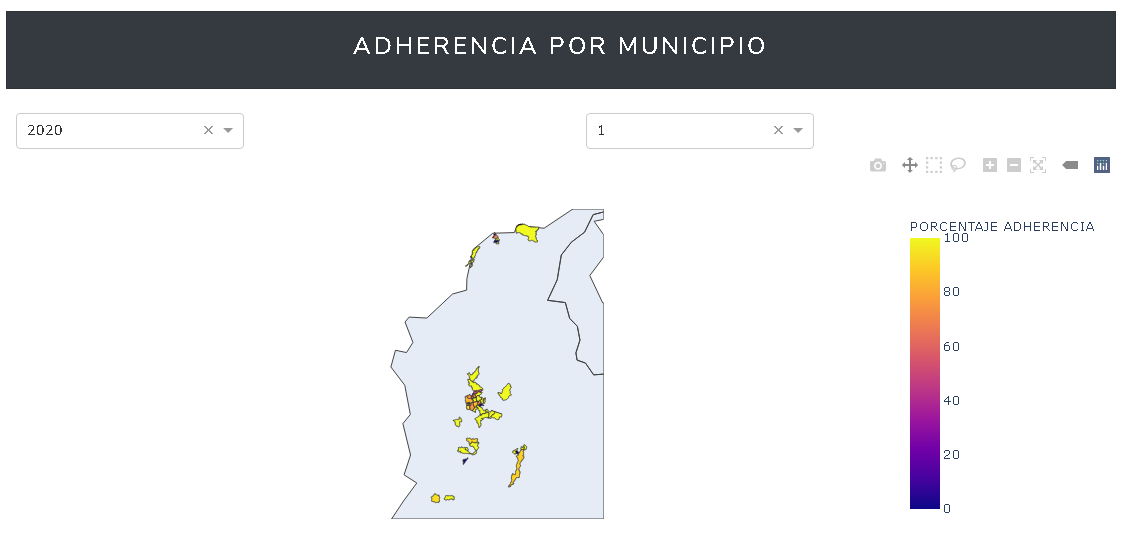
Figure XXVII: *Module 1 (part 3)*.

**

1. Modulo\_2.py is composed of a characterization of how a patient’s adherence evolves over time and geo-localization.

First, you can find a map with some Colombian regions from which we have data regarding the adherence of patients over time. For this map, you can filter the data by year and month. Concerning spatial distribution, it was constructed based on the division of municipalities and/or districts according to the codification of municipalities proposed by the national statistics bureau (DIVIPOLA - DANE). Finally, concerning the map, you can find the color scale of the level of adherence that there is for each of the registered municipalities.

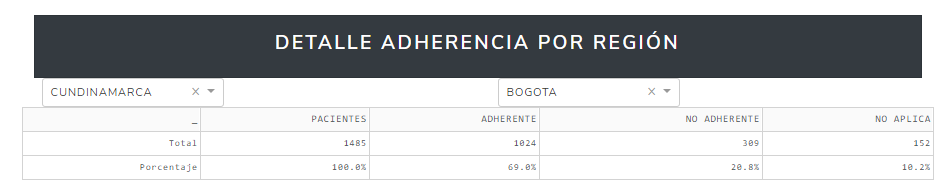
Figure XXVIII: *Module 2 (part 1)*.



Secondly, we have the number of patients reported by a municipality, the adherent,, and non-adherent numbers. Also, the frequency presented for each of these values.

Following this, the risk factors associated with non-adherence by patients can be found. These risk factors can be filtered out based on spatial criteria, such as a municipality, or temporal criteria such as year and month.

Figure XXIX: *Module 2 (part 2)*.



1. RDS\_funciones.py contains all the necessary functions to be able to do the queries between the databases loaded in AWS and relate them to the modules described.

For the health professionals module, you can see the chart to enter the patient's identification number which will serve as a key for the back office consultations and which will bring 1) the patient's demographic information, 2) the evolution of medication claims to time and 3) the calculated probability of adherence for that particular patient.

# Conclusions

Adherence of treatment is vital in the risk reduction for patients in chronic diseases. Chronic diseases normally do not have a cure, but just a mild treatment that allows a person to have a generally better quality of life than the expected for the condition. This is the asthma case, where the respiratory functions are compromised and trigger factors are found in the general environment of any citizen.

Particularly, when treating severe asthma patients, there is a majority of adult and elderly population that tend to have other relevant comorbidities. These combination of conditions result in the fact that this group tends to have a high rate of adherence (79%).

Constructing different models to estimate the probability of adherence, we observed that tree-based machine learning methodologies presented better performance than traditional models, and these methodologies do not require rigorous compliance of statistical assumptions.

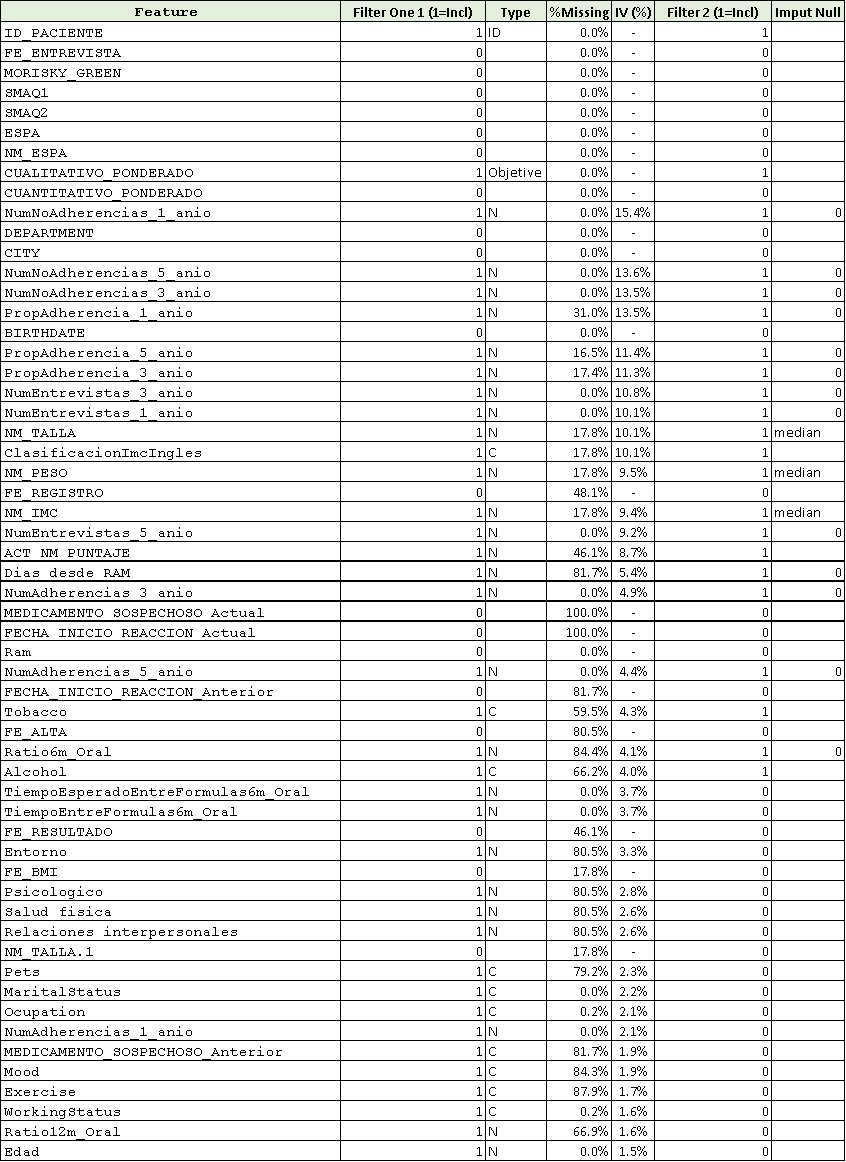
Since adherence was the main objective for modeling, a second model based on persistence in claiming medications were built . This provides an opportunity to generate additional tools in order to establish strategies to have a better health risk management.

A dashboard was developed to analyze many factors of patients related to their health. Data that were later used to calculate the patient's possible adherence to medications. Several characteristics were added, taking into account the needs of the Omnivida company, all of them aimed at creating policies and procedures to increase patient adherence nationwide.

The proposal is to have a welcome (landing) page for the user, where various indications and descriptions of the two modules that can be accessed are given. The first module shows specific characteristics of the patients, while module two uses, for example, maps to present the current situation of patient adherence at the national level to the user.

Annexed A

List of covariates.





1. Ministerio de Salud y Protección Social, “Cubos sispro: Registros Individuales de Prestaciones en Salud”. Accessed: August 1st, 2020. [↑](#footnote-ref-0)
2. Departamento Administrativo Nacional de Estadística, Ministerio de Salud y Protección Social, “Cubos Sispro: Estadísticas Vitales (Defunciones)”. Accessed: August 1st, 2020 [↑](#footnote-ref-1)
3. Ministerio de Salud y Protección Social, “Cubos sispro: Registros Individuales de Prestaciones en Salud”. Accessed: August 1st, 2020. [↑](#footnote-ref-2)
4. Ministerio de Salud y Protección Social, “Cubos sispro: Registros Individuales de Prestaciones en Salud”. Accessed: August 1st, 2020. [↑](#footnote-ref-3)
5. Departamento Administrativo Nacional de Estadística, “Proyección Poblacional (2018-2023) Anexo: Departamentos”. Accessed: August 1st, 2020 [↑](#footnote-ref-4)