NOVA SCHOOL OF SCIENCE & TECHNOLOGY

MyBreath - Manage Your Breath

SAMSUNG AI course 2022 @ NOVA

Report of the final project

Group:

Amin Khodamoradi

Ana Carolina Pádua

Joana Seringa

Zabya Abo Aljadayel

1. Introduction

Asthma is a long-term condition, affecting 339 million people worldwide (The Global Asthma Report 2018 | The Union), often with seasonal and lifetime differences in symptoms and disease burden. Although, asthma symptoms can be controlled most of the time, some have on-going poor control and all are at risk of attacks which, at best, are inconvenient and at worst can result in hospitalisation or even death (An Official American Thoracic Society/European Respiratory Society Statement: Asthma Control and Exacerbations | Standardizing Endpoints for Clinical Asthma Trials and Clinical Practice | American Journal of Respiratory and Critical Care Medicine (atsjournals.org)). Currently, there is no cure for asthma, therefore it is a chronic condition, whose focus of management is on improving symptom control and reducing the risk of attacks.

To contribute to the better management of this chronic condition, we developed this project whose main objective was to create a machine learning model to predict the level of risk to have an emergency hospital admissions due to asthma attacks, considering environment conditions (weather, pollen and pollution).

Although there are some apps and machine learning models for asthma management (https://www.dovepress.com/getfile.php?fileID=81792), to the best of our knowledge this is the first project that considers all the environment conditions mentioned above. The development and improvement of this model is intended to be used, in the future, to provide an alert system to healthcare professionals and people with asthma considering two levels of risk (low and high risk). This alert system could be deployed by an app and integrated in the official system information used in hospitals. Therefore, there would be two customer segments: people with asthma (that could buy the app or have access to a free version that would not give personalised information) and healthcare institutions, whose revenue structure would lay in a commercial contract to implement and maintain the system.

Considering this is a tool to support decision-making, the preferred channels to communicate this product are conferences and scientific societies, through healthcare professionals. Therefore, the cost structure of this system would include development costs, maintenance costs and marketing costs.

In the following chapters we are going to present the methodology used to develop the machine learning model, as well as the discussion and conclusions.

2. Methods

The methodology of this project followed the main steps to build a machine learning model: (1) data collection; (2) data preparation; (3) model selection and implementation; (4) model evaluation. Following we are going to detail each of these steps. All the code was developed with Python and R, using Jupyter notebook.

2.1. Data Collection and Integration

Regarding the data collection, we collected and integrated the following data:

- Admission data (from a hospital in the Area of Greater Lisbon);
- Weather data (Historical Weather Data & Weather Forecast Data | Visual Crossing);
- Pollen data (RPA - Rede Portuguesa de Aerobiologia (rpaerobiologia.com));
- Pollution data (Air Quality Historical Data Platform (aqicn.org)).

Admission Data

The admission dataset was from one hospital in the area of Greater Lisbon. Greater Lisbon is the region surrounding Lisbon, the capital of Portugal. The region includes a variety of localities and it is the main economical subregion of the country. It is the most populous and most densely populated Portuguese subregion.

The dataset (*URG_ASMA_LAST_with_2022.xlsx*) had a total of 1,762 episodes of emergency asthma admissions during the period in analysis, which is from 01 January 2019 to 29 November 2022. We excluded data after March 2020 until 2022, because during the months of COVID-2019 there was a significant decrease in asthma admissions (related to the fact that individuals were more isolated). This dataset contains demographic characteristics of the patients such as sex, age, diagnosis and data of admission to the emergency department of the hospital.

Pollen data

Specifically regarding pollen data we created a dedicated script that used web scraping for pulling data out of the website regarding daily pollen intensity levels.

The script file named *Phase 1, read pollen data.ipynb* reads several links by updating one base link with the different archives available for the year of 2019 and 2022. There was no pollen data regarding the year 2020. This was implemented using the funcion *replace*.

The pollen intensity in the region of Lisboa and Setubal was selected for each one of the archives. For that, the function *find_all()* from *beautifulsoup* was used. We took advantage of the fact that the pollen intensity was mentioned in bold in the body text of the website to extract that information, using this text feature. The pollen data file is named Pollen_2019_2022.xlsx.

Weather and Pollution data

Regarding weather and pollution data we extracted the respective excel and csv files directly from the websites.

Weather data is called *Lisbon,Portugal 2019-01-01 to 2022-11-29.xlsx* and pollution data is named *entrecampos_lisboa-air-quality.csv.*

Data from admissions, pollen, weather and pollution was then integrated (by *Phase 2, Data integration.ipynb*) in a single file, based on date (weather), named *integrated Asma Data(weather, admission, pollen, pollution).csv*.

2.2 Data Preparation

2.2.1 Data Cleaning and Exploratory Data Analysis

After the data collection and integration, we did data preparation (by *Phase 3, data cleaning and visualization.ipynb*), where we excluded variables with more than 30% of missing values *(‘conditions’, ‘windgust’, ‘preciptype’, ‘severerisk’, ‘sunrise’, 'moonphase’, ‘sunset’, ‘icon’, ‘description’).*

Regarding pollution data we replaced the missing values with zero (0), once according to the website the null values were zeros.

We also deleted other variables (*'data_admissao', 'dicofre', 'distrito', 'concelho', 'freguesia', 'regiao', 'nacionalidade', 'cod_proveniencia', 'proveniencia', 'cod_causa', 'causa', 'cod_destino', 'destino', 'diagnostico', 'classe', 'asma', 'date', 'name','snow', 'snowdepth', 'winddir', 'cor', 'stations')*, because they were not relevant for the problem under analysis or had the exact value for all the episodes.

The cleaned data file is named *'daily data.csv'*. This data includes 789 rows and 24 features and “admission number” as target column.

After cleaning, we did statistics and visualisation. We started by checking the distribution of daily pollen intensity (table 1) and the daily distribution of the number of admissions (table 2).

Level of Pollen Intensity	Number of days
good	565
baixos	28
moderados	45
elevados	28
muito elevados	123

Table 1. Daily Pollen Intensity

Number of Admissions	Number of days
0	126
1	183
2	180
3	133
4	82
5	45
6	24
7	9
8	3
9	4

Table 2. Daily distribution of the number of admissions

We defined the risk level (low risk and high risk) based on the number of admissions per day, where low risk was inferior or equal to two admissions per day, and high risk equal to three or more admissions per day.

Later, based on daily data, we did a correlation matrix (image 1) in order to find linear correlation between our target (“risk level”) and features.

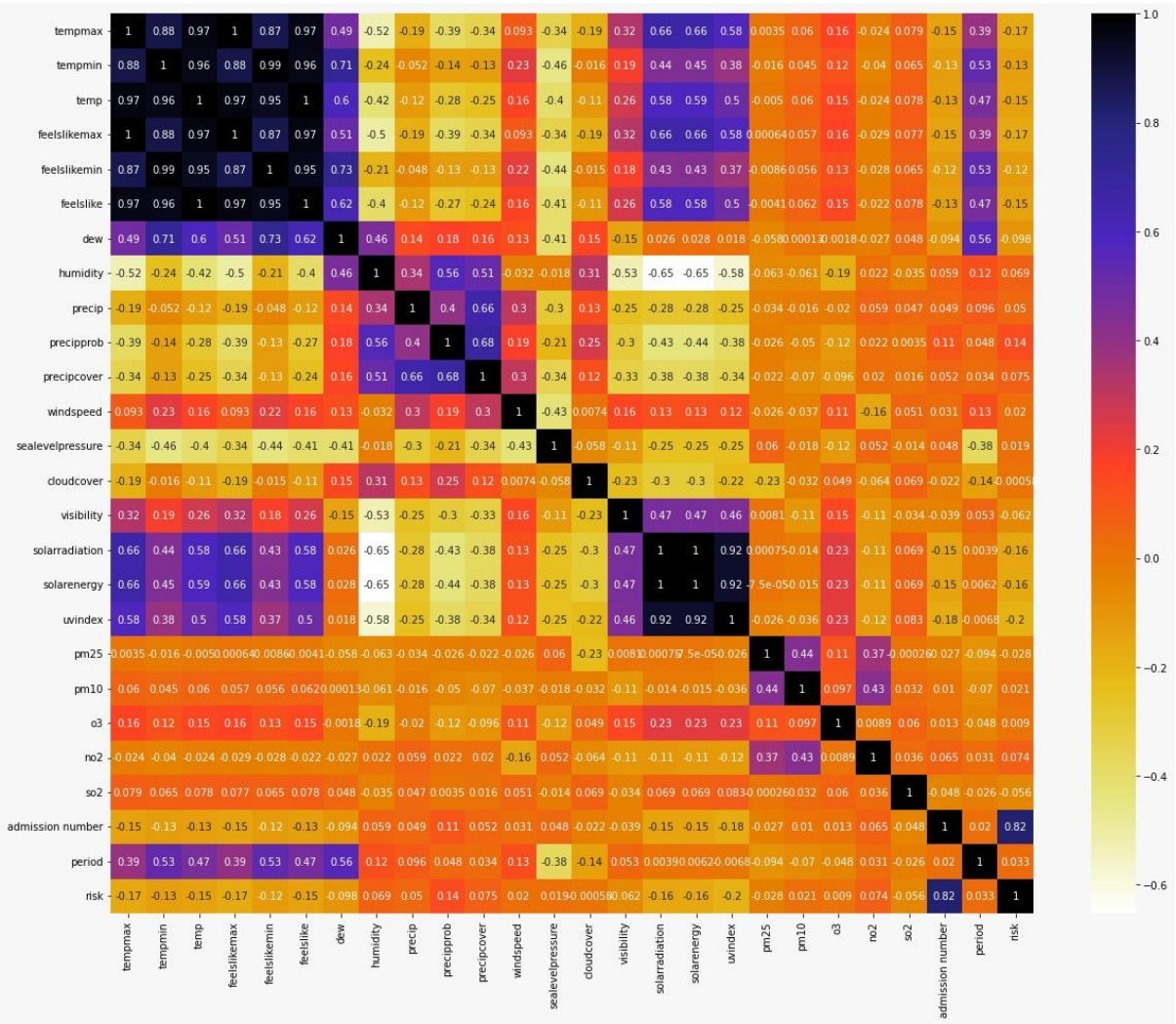

Image 1. Correlation matrix

After that analysis we checked the number of admissions per day of the week and month (image 2), as well as the risk level per day of the week and month, where dark green shows high risk and light green shows low risk.

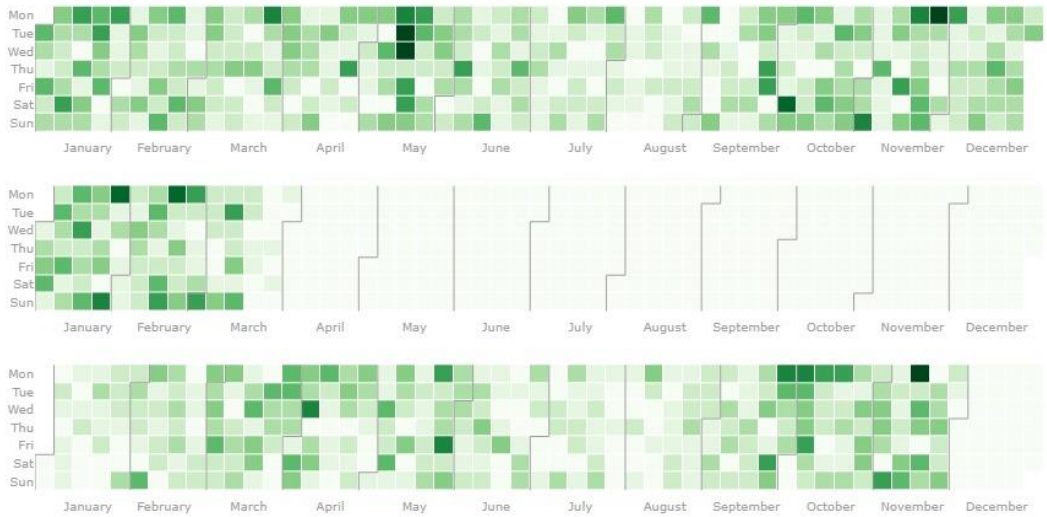

Image 2. Number of admissions per day of the week and month

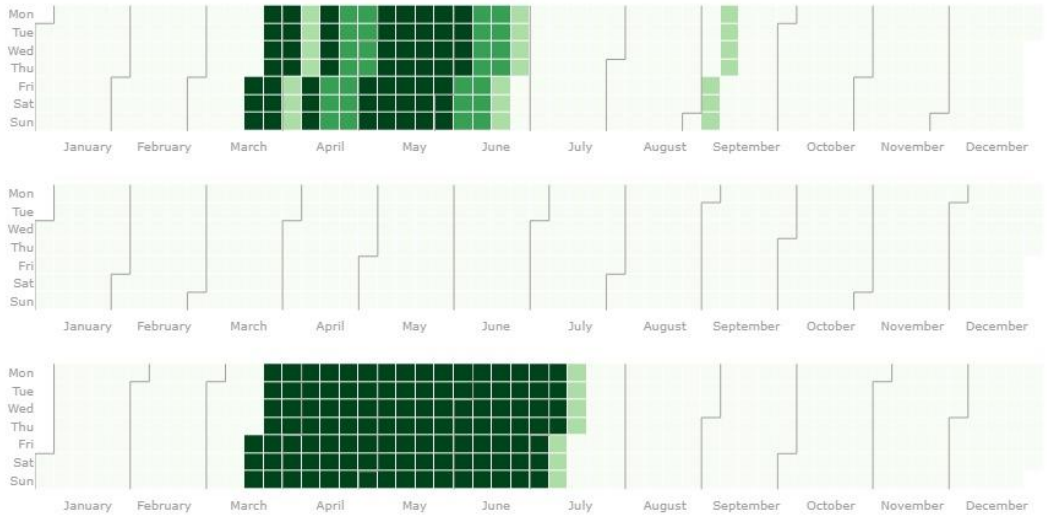

Image 3. Daily risk level

Image 4 plots the admission number, showing the median equal to 2 admissions per day.

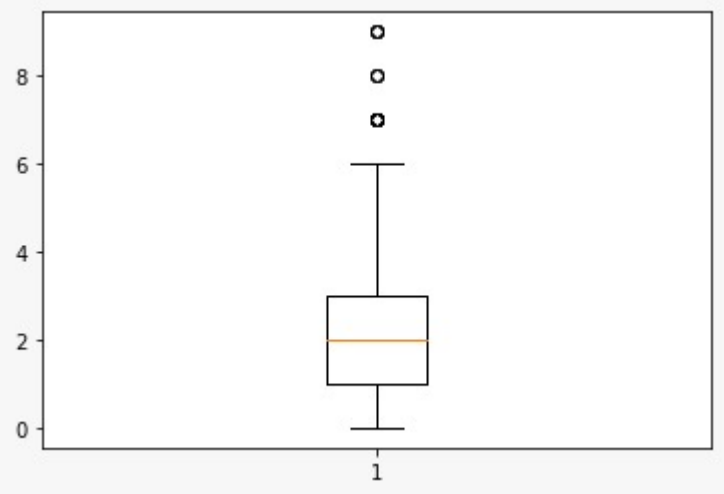

Image 4. Admission number box plot

We also visualised the feature distribution (image 5).

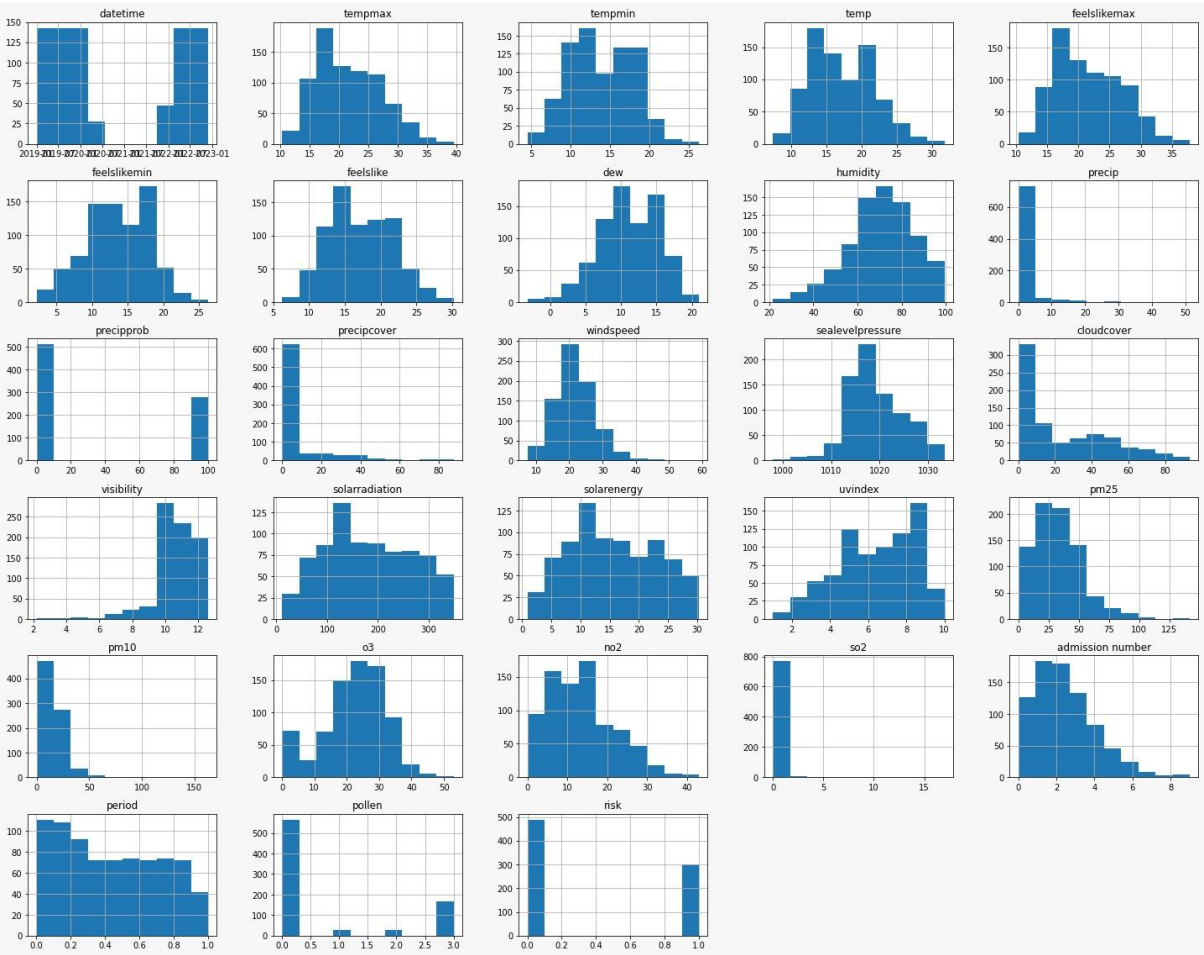

Image 5. Distribution for all the features

2.2.2 Feature Engineering

After data cleaning and visualisation, we started with feature engineering. The file related to this step is called *Phase 4.1, Classification.ipynb*.

Firstly, in the preprocessing we converted categorical features to numerical features, and created features based on previous ones (such as “temp amplitude”, “temp change 1” and “temp change 2”).

Train and test split: as our data is time-series, we divided our data to train and test in which 80 % was for training and 20 % for test. In the following, until testing our models we always did the feature selection and model tuning only on train data.

Then, for feature selection we tried a set of feature selection methods:

- **Stepwise feature selection** (backward and forward), implemented on R (file named *Stepwise feature selection(backward and forward).R*);
- **Chi-square test**, which is a statistical hypothesis test used when the sample sizes are large to examine whether two categorical variables are independent in influencing the test statistic (Chi-Square - Sociology 3112 - Department of Sociology - The University of utah). We evaluated that based on the F_score and p value; The threshold applied for p-value was <0.16.
- **Mutual Information**, which is a quantity that measures a relationship between two random variables that are sampled simultaneously, and that showed no linear correlation;
- **L1-based feature selection,** which removes features with low variance;
- **Extra tree feature importance** to decide which variables are important to be included in the model (image 6).

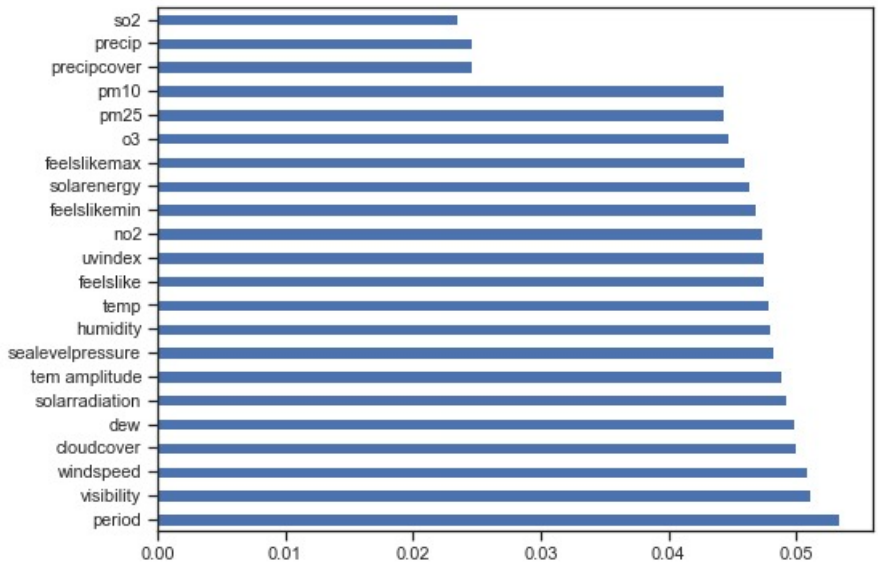

Image 6. Extra tree feature importance

After trying all the methods mentioned above we selected the Chi-square test with criteria of p-value < 0.05.

The variables included in the model were:

1. Precipprob;
2. Solarradiation;
3. Precipcover;
4. Solarenergy;
5. Precip;
6. So2;
7. Pm25;
8. Uvindex;
9. Feelslikemax;
10. No2;
11. Feelslike;
12. Humidity;
13. Temp;
14. Feelslikemin;
15. Tem amplitude

2.3 Model Tuning

Once our dataset is based on data samples, we used GridSearch specialised for Time Series data as follows: train and test TimeSeriesSplit, the Time Series cross-validator, with 3 splits. Time-series split is a special kind of train-test split. The object for the time series split is similar to random split which is to validate the model predictability regardless of how train-test data sets are split. However, the time series split ensures, in this case, the test datasets are younger than train datasets, which is more realistic since we will not be able to train on “future” data.

We tried a set of machine learning algorithms to understand each one had a better predictive power, which were:

- Linear Regression
- Logistic Regression
- Linear Discriminant Analysis
- Classification and Regression Trees
- Gaussian Naive Bayes
- Support Vector Machines (SVM)
- Random Forest
- Sklearn Gradient Boosting
- XGBoost Classifier
- XGBoost Random Forest Classifier

2.4 Model Evaluation

Model evaluation concerns the test data (the recent 20% of the dataset). Additionally to the accuracy of the algorithm, we used Precision, which answers the question: “What proportion of positive identifications was actually correct?”, Recall (also known as sensitivity), which answers the question: “What proportion of actual positives was identified correctly?”, F1-score, which combines the precision and recall of a classifier into a single metric by taking their harmonic mean, and Mean Squared Error (MSE), which measures the average squared difference between the estimated values and the actual value. Table 3 presents methods of evaluation according to F1-score weighted and MSE.

	method name	F1-score weghted	MSE
0	Gaussian Naive Bayes	0.62	0.316
1	Linear Discriminant Analysis	0.68	0.291
2	Decision Tree	0.66	0.335
3	SVM	0.66	0.31
4	Logistic Regression	0.67	0.297
5	Sklearn-Gradient Boosting	0.62	0.367
6	XGBoost Classifier	0.6	0.386
7	XGBoost random forest	0.59	0.335
8	Random Forest model	0.63	0.341

Table 3. Methods evaluation

3. Discussion

3.1. Limitations

This project had several limitations also reported in other projects, such as data quality once real-world settings are more likely to lead to reduced data quality. Other limitations are related to the (low) sample size related to the fact that we only used information from one hospital and just for one year and three months. Another limitation was that data available from the period of april 2020 to december 2021, inclusive, could not be used, because the number of hospital admissions was highly influenced by the COVID-19 pandemic.

3.2. Future Direction

Future projects should consider more personal characteristics, not only weather characteristics because asthma attacks depend on personal characteristics. Also, to achieve the future goal to use this model to provide an alert system to healthcare professionals and people with asthma considering two levels of risk (low and high risk) it will be needed to work with information from all the country.

4. Conclusions

The data analysed does not suggest a direct influence of environment conditions (weather, pollution and pollen) in the risk of hospital admissions due to asthma.
This might be due to the fact that the dataset used is limited. The number of samples should be increased in order to allow taking conclusions. For instance, in a further analysis we would use data from several years and not focus on a single hospital.

Studies indicate that this risk is highly influenced by personal characteristics. Therefore, in the second part of this study, we would explore the impact of the individual characteristics (sex, age, existence of other diseases) on the number of asthma admissions at the hospital to give more personalized recommendations.