Miniproject M1: Security risk on a smart home

Almudena Maroto

January 2021

Contents

1 Abstract

The Internet of Things (IoT) has arrived to change the life of humanity. Its purpose is connecting every electronic device used. Thus, giving the user the ability to control every electronic item in his surroundings, from turning on the TV to locking doors or changing heating temperature. Everything is connected through internet into a single controller, giving the user a total control of his home.

Smart homes are designed to provide a better quality life to the user and even to promote a responsible energy consumption. According to the results obtained by the Schlage's Industry Insight Survey [1], the 86 % of the young population would be willing to pay more rent to live in a smart house. Even the 44 % of them would exchange a parking lot for living in a smart home. Therefore, Smart Homes promises to have a bright future.

Although, IoT brings great benefits it can not be forgotten the possibles risks that comes associated to it. The data stored from the devices contains crucial information about the users. The stored data from the electronic devices can be used to detect patterns in the behaviour and suppose a risk to the security. This make the user vulnerable to cyber attacks, that could derive on physical robberies in home users.

The purpose of the present project is to analyse the limits of security in smart homes, and discover how easily could be to identify the patterns of behaviour of a smart home user. This project wants to rise a warning about the risk that can bring an inadequate use of the stored data.

1.1 Performance of a smart home

The smart homes analysed in this project counts with several sensors of all kind all around the house. The sensors include:

- Electricity: Power consumption, current, tension and power factor.
- Temperature, humidity and brightness.
- Aperture of windows and doors.
- State of the lightening.
- Analysis of the quality of the air.
- Atmospheric conditions in the weather station.

The data used in this project comes from the Expe-smarthouse project [2]. It provides data from a $120m^2$ household where five person lives. The analysis will be focused on the information registered during November 2020.

With the objective of mark the importance of security in smart homes data, this project will pretend to detect patterns in the behaviour of the users, and discover what useful information could be found for a thief if he could have access to the data. Therefore, the purpose of this data analysis will be to detect **when the house is empty** making it a good moment to breach the threshold of the property. It will be employed conventional data analysis as well as a Decision Tree Classifier.

Key words : Data Analysis, Smart Homes, Decision Tree.

2 Data collection

The first step will be collect the data. The data is downloaded from the Grafana website [3]. It is obtained in *csv* format, and will be treated with Python. In the appendix can be found the code used with this purpose.

It is important to remark that the data proceeds from different files. Therefore, it needs to be joined in a way that each row of the data contains the instant value of each sensor at certain date time.

	Time	Hum_Cuisine	Hum_Extérieur	Hum_2ème étage	Hum_Chambre 3	...	Year	Month_name	Weekday	Day_number	Hour
0	2020-10-26 00:00:00	66.0	82.0	77.0	71.0	...	2020	October	Monday	26	0
1	2020-10-26 01:00:00	64.0	85.0	77.0	70.0	...	2020	October	Monday	26	1
2	2020-10-26 02:00:00	63.0	84.0	76.0	70.0	...	2020	October	Monday	26	2
3	2020-10-26 03:00:00	61.0	81.0	75.0	69.0	...	2020	October	Monday	26	3
4	2020-10-26 04:00:00	61.0	80.0	75.0	69.0	...	2020	October	Monday	26	4
...	...	...	...	...	...	...	...	...	...	...	...
763	2020-11-26 19:00:00	49.0	74.0	61.0	54.0	...	2020	November	Thursday	26	19
764	2020-11-26 20:00:00	47.0	75.0	62.0	55.0	...	2020	November	Thursday	26	20
765	2020-11-26 21:00:00	47.0	75.0	62.0	56.0	...	2020	November	Thursday	26	21
766	2020-11-26 22:00:00	47.0	77.0	61.0	56.0	...	2020	November	Thursday	26	22
767	2020-11-26 23:00:00	46.0	77.0	62.0	57.0	...	2020	November	Thursday	26	23

768 rows × 86 columns

Figure 1: Data set structure

Besides, some files contain different frequency of collection data. For example, one sensor can collect the data daily, while another sensor is taking data hourly. To fix this problem, it has been interpolated the data in those sensors that needed a higher sampling frequency. For the data set with higher frequency that used, it has been taking the hourly mean value.

Another needed treatment into the data is to manage the outliers. The outliers are certain data points that have been collected incorrectly and have higher values than normal. Meaning that if they are kept into the data set, they would distort the results, and no valuable information could be obtained. The solution to this problem is to delete the data points that have a value higher than the mean value of the data plus three standard deviations. As it can be seen in Figure 2, this condition deletes around the 0.3% of the data points.

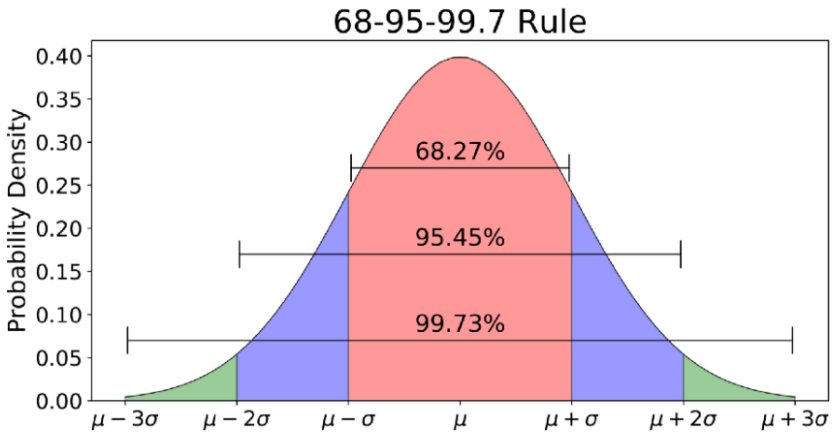

Figure 2: Detection of data points outside three standard deviation [4]

3 Analysis of sensors data

In this section, it will be analysed different sensors to see if the stored data from the sensor can be used to detect pattern behaviour of the user. It will be analysed different variables, taking special consideration on power demand and air quality.

3.1 Power demand

The smart home counts with several sensors connected to electric devices, including TV, washing machine, dishwasher, oven among others. Since the data has hourly frequency, it is possible to compute the mean power demand at each hour as well as at each day. In Figure 3, it can be seen the average power consumed by each electronic device, as well, as the total power demand.

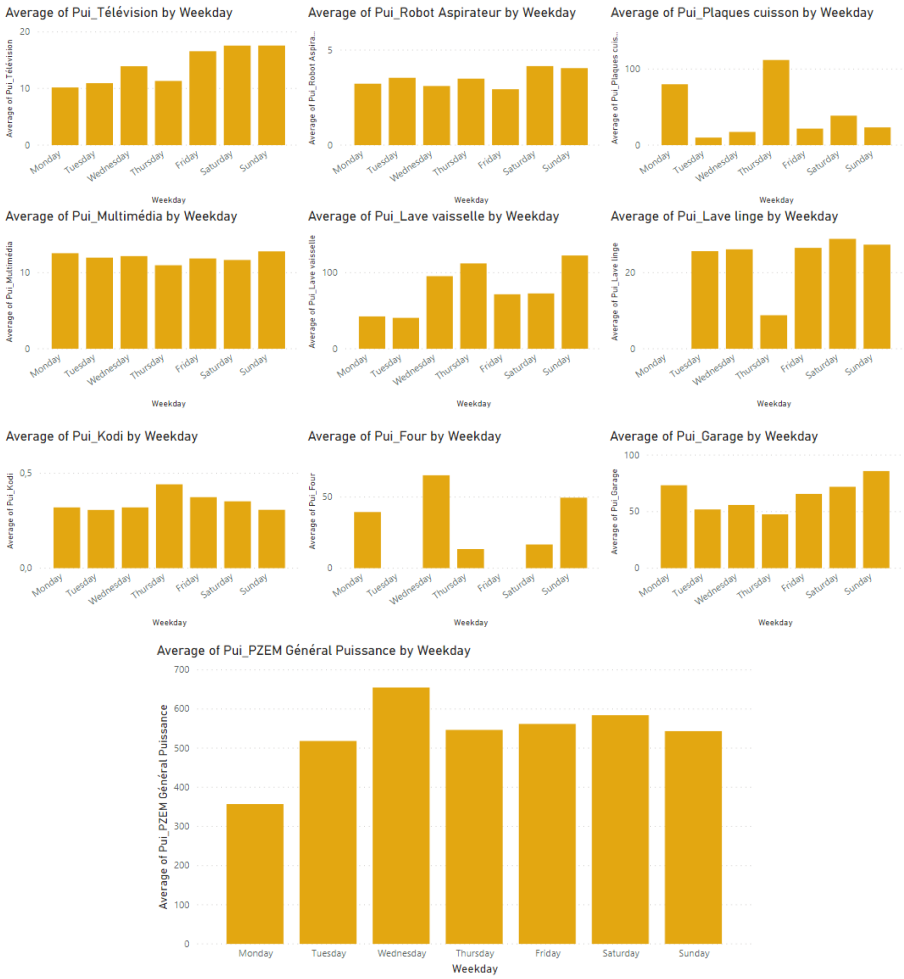

Figure 3: Variation of power demand over the week

The first thing to notice is the different consumes of energy that have the electronic devices. The hotplates, dishwasher and garage are the devices with higher consumption. Although, the hotplates are not commonly used every day, and the user shows a preference to cook on Mondays and Thursdays. Other devices seems to be connected or used every day without any preference,

as the robot vacuum cleaner and multimedia. In general, Mondays are days with less consume at home, which could mean that the habitants are not home.

It is also useful to define at which hour are the devices being used. Figure 4 shows the power demand in each hour of the day.

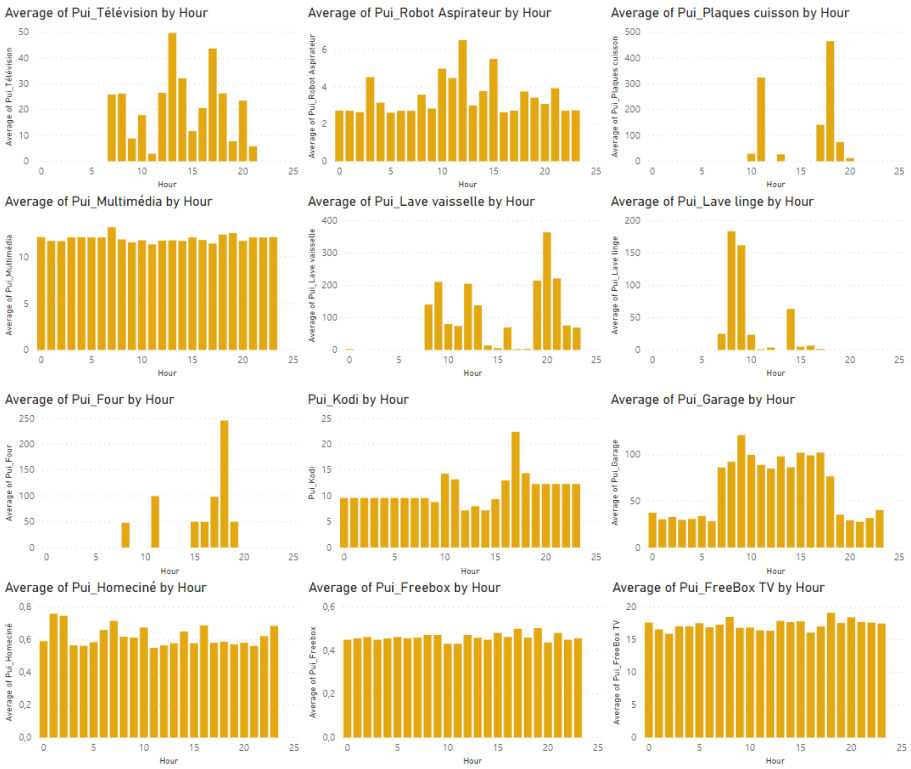

Figure 4: Variation of power demand at each hour

As it happened before, there are devices that are consuming a constant energy every day and every hour of the week. But now, it is possible to concrete at which moment of the day are used some electric devices: The hotplates are used mostly at 11 AM and 18 AM, signaling the preferred hour of the user to be in the kitchen. The TV use shows certain hours of preferred use, and can be use to define the hour in which the user wakes up (7 AM), and the hour in which it turns off the TV, probably to go to sleep (20 - 21 AM). In Figure 5 can be seen the total power consumption.

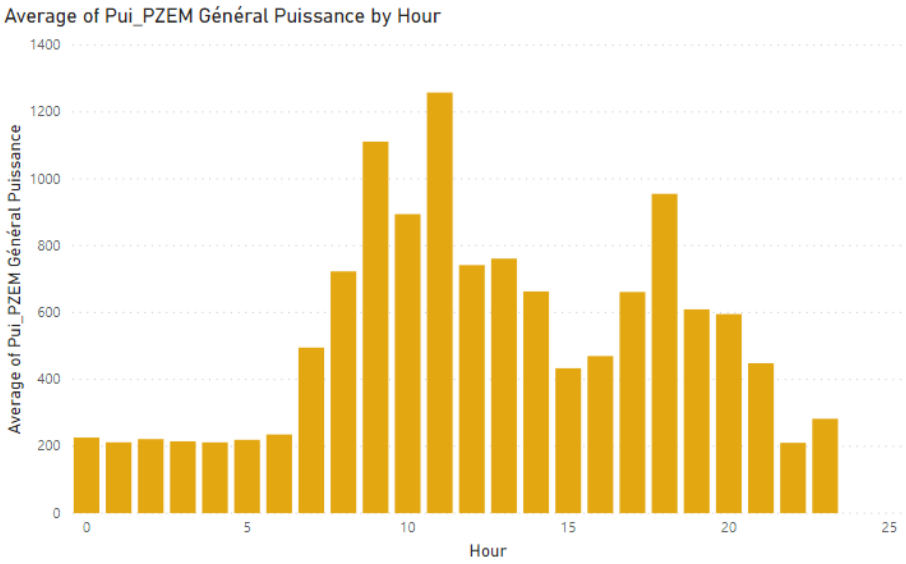

Figure 5: Variation of total power demand at each hour

Although, since it was defined that Monday is probably the day when none is at home, it would be better to focus only in the consume in that day. Figure 6 shows the consume that specific day.

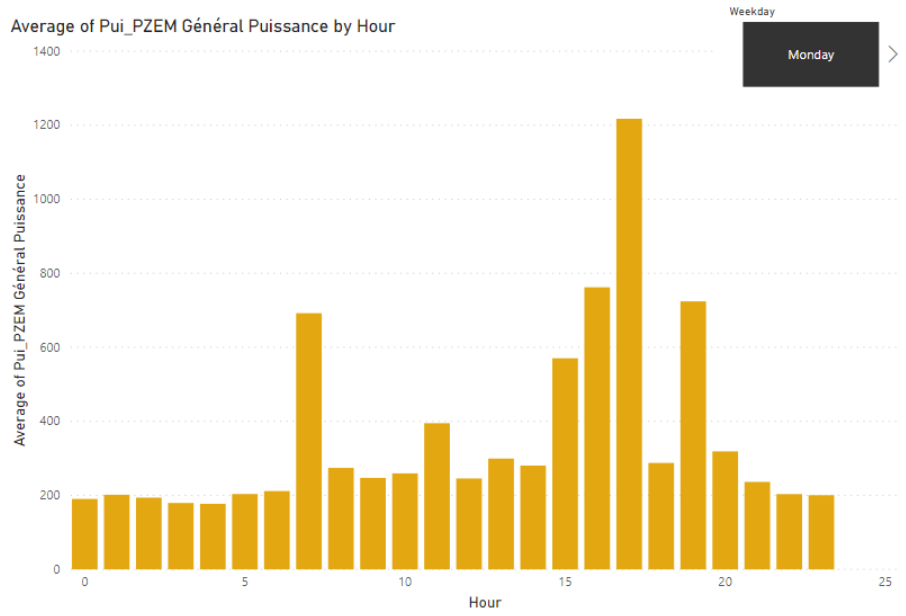

Figure 6: Variation of total power demand at each hour

The peak hours of consume in Mondays are in the afternoon, with a small peak at 7 AM. With only this data there are two possible options: the user works at night and usually sleep in the mornings, or the user works in the morning and arrive home around 3 PM. Therefore, to discover what is the real option it is needed to include more variables into the analysis.

3.2 Air quality analysis

The measurement of CO_2 concentration in a chamber is a key indicator in the occupancy of a room. The higher the concentration, the longer time has been a person in a room. Figures 7 and 8 shows the evolution of the concentration of CO_2 in the different chamber of the house.

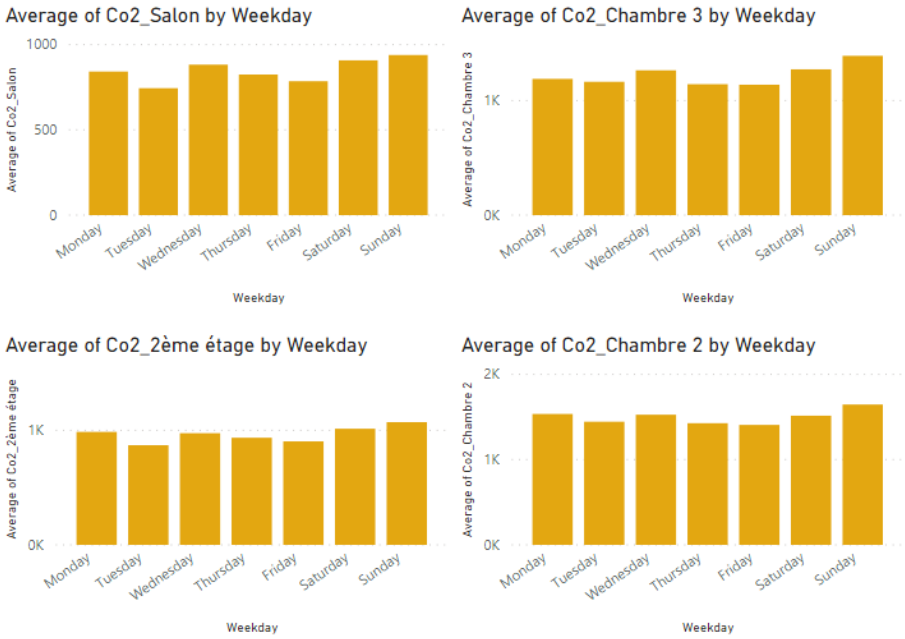

Figure 7: Daily variation of CO_2 concentration

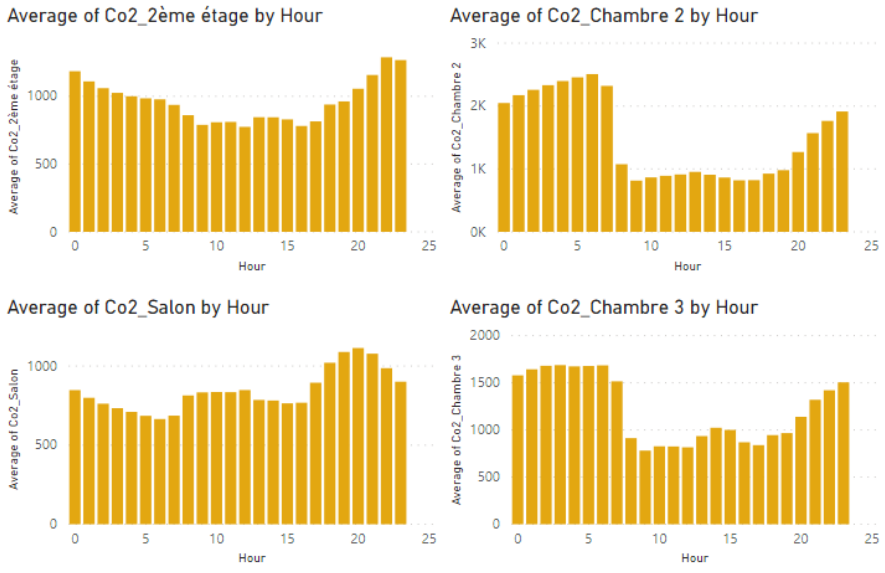

Figure 8: Hourly variation of CO_2 concentration

The analysis of the results shows small difference in the CO_2 concentration along the week. It only

shows a small higher value on the weekends, but not enough relevant to be considered. However, the hourly analysis shows bigger differences. The CO_2 decrease in the daily hours while it increases in the nights.

Since the concentration is a indicator of occupancy in the room, the increase at night could mean that the user is at home sleeping. The peaks of concentration show the movement along the rooms. During the afternoon the user can be found in the living room, and thereafter he moves to the chambers. The concentration in the rooms keeps high until the 7 AM. At this moment, as it was shown in the power demand analysis, the user wakes up.

4 Correlation between sensors

Another point of view that could be to analyse if the weather conditions has any influence on the pattern behaviour of the user. The Figure 9 serves to discover if there is any correlation between the weather and some sensors data that are related to the user behaviour.

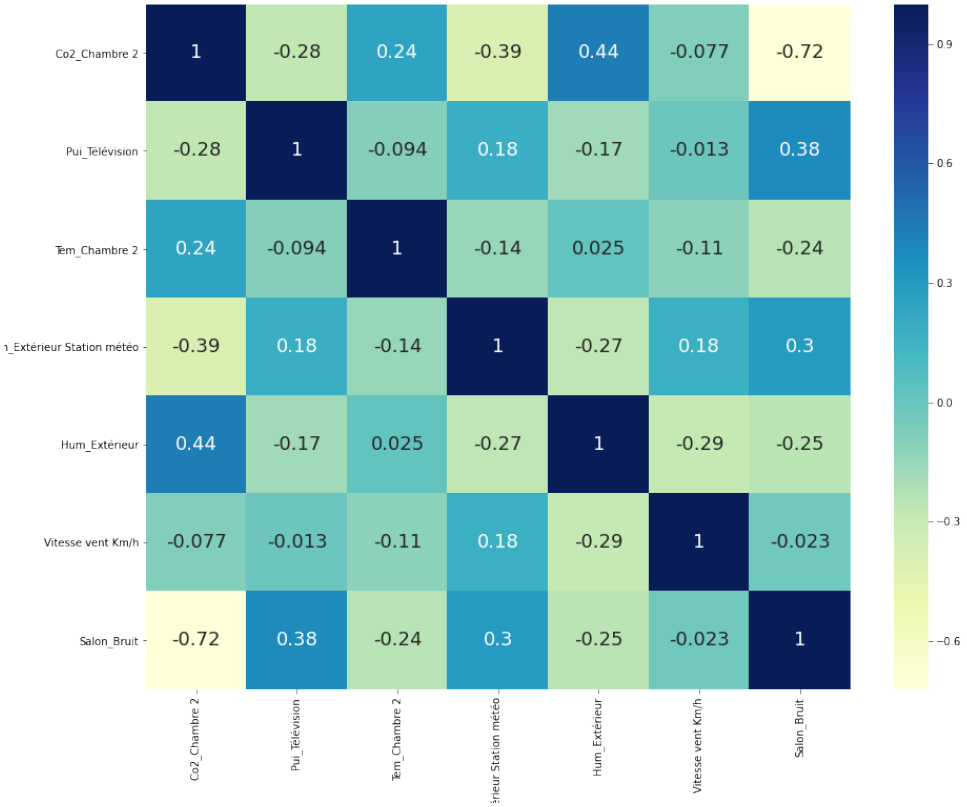

Figure 9: Correlation between atmospheric conditions and occupancy of the house

The correlation matrix shows the relationship between two variables. In the case of the CO_2 in the chamber, there is a negative correlation with the power consumed by the TV and the temperature atmospheric. While it has a positive correlation with the temperature of the chamber, the exterior humidity and the noise in the living room. This correlation makes sense, since when the user is

in the living room, the CO_2 in the chamber will decrease. On the contrary, the temperature of the chamber will be higher if the habitant is inside, and therefore the CO_2 concentration will be higher too. However, there are other variables without a significant correlation with the air quality of the room, as it is the wind speed.

Regarding the power consume on TV, it has lower correlation with the weather conditions. This could mean that when the weather get worse, the user has a preference of the chamber, instead of the living room. However, the most probable meaning is that the weather conditions are worse at night, and it is in that moment when the user rest on the chamber.

In resume, there are too many variables in play to be able to guess the occupancy of the house by visual analysis of the results. Luckily, it exists more advanced analysis techniques that can help to manage great amounts of variables.

5 Decision Tree Classifier

In this section, it is proposed a Decision Tree Classifier to find the occupancy of the house. This technique is one of the most used algorithms in machine learning. It is an supervised method technique where the model needs a target variable to deliver results. Its structure is formed by nodes, where certain variable is measured. Depending on the value of the variable, the flowchart will follow one of the branches. The Decision Tree are easily constructed and interpreted and it only uses the most important variables.

The first step to apply it in the data set is to define the target variable. Since the data set that is treated in this project does not contain an occupancy variable, it will be employed the *Entry Movement* variable. This sensor receives a signal when there is a movement in the entry of the house. If there is a person inside the house without moving (for example in another chamber) it will not detect the occupancy of the house. However, due to the lack of a more accurate variable for the purpose of this project, it will be employed as such.

The rest of the variables will be used as input data. Although there is a consideration to be made: The data set counts with categorical features, as the day of the week or the month. The Python library that will be used to build this model only accepts continuous data as input. Therefore, it will be needed a one-hot-encoding to convert the categorical features into continuous numerical data.

One-hot-encoding consists on making n new columns in the data set, each of them with one value of the categorical data. For example, if the week of the day is one-hot-encoding, it will be converted on 7 columns. Each column will have a one if it is that day of the week or a zero if not.

	Time	Weekday	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
0	2020-10-26 00:00:00	Monday	0	1	0	0	0	0	0
1	2020-10-26 01:00:00	Monday	0	1	0	0	0	0	0
2	2020-10-26 02:00:00	Monday	0	1	0	0	0	0	0
3	2020-10-26 03:00:00	Monday	0	1	0	0	0	0	0
4	2020-10-26 04:00:00	Monday	0	1	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
763	2020-11-26 19:00:00	Thursday	0	0	0	0	1	0	0
764	2020-11-26 20:00:00	Thursday	0	0	0	0	1	0	0
765	2020-11-26 21:00:00	Thursday	0	0	0	0	1	0	0
766	2020-11-26 22:00:00	Thursday	0	0	0	0	1	0	0
767	2020-11-26 23:00:00	Thursday	0	0	0	0	1	0	0

Figure 10: One-hot-encoding example

Besides, the data set will be divided into train and test data to be able to compute the accuracy of the model and how it behaves with unseen data. The data is divided randomly: a certain percentages of the rows data chosen for the test set. With all that into consideration, it is obtained the Decision Tree from the Figure 11. The model has an accuracy of 88.0 %.

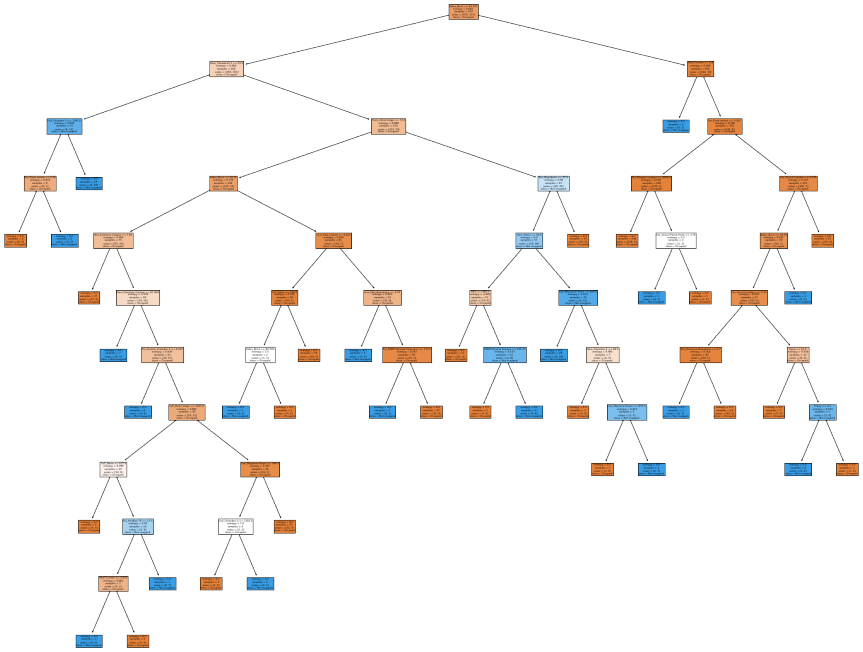

Figure 11: Structure Decision Tree

6 Conclusions

To summarize, these are the conclusions that can be extracted from the present project:

- The smart homes can be a good ally to help the user to made a more efficient use of his energy consume, even make his life easier. However, they have a inherent risk that must be known by the user to be treated and protected properly.
- Machine learning can reach conclusions that are hard to achieve with the conventional data analysis.

References

[1] Shclage. *Results of Schlage's Industry Insight Survey Reveals What Millennial Renters Want in 2017*. 2016. URL: https://www.prnewswire.com/news-releases/results-of-schlages-industry-insight-survey-reveals-what-millennial-renters-want-in-2017-300369197.html.

[2] G scop. *EXPE-SMARTHOUSE: The full connected living house*. 2018. URL: http://expe-smarthouse.duckdns.org/?page_id=7&lang=en.

[3] Jerome Ferrari. *Grafana data home*. 2020. URL: https://jarvis-oneforall.duckdns.org:3000.

[4] Michael Galarnyk. *Explaining the 68-95-99.7 rule for a Normal Distribution*. 2018. URL: https://towardsdatascience.com/understanding-the-68-95-99-7-rule-for-a-normal-distribution-b7b7cbf760c2.