INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Energy Consumption Forecasting – A Proposed Framework

Hugo Miguel Nogueira Mendes

A Dissertation presented in partial fulfilment of the requirements of the Degree of **Master in Integrated Business Intelligence Systems**

Supervisors:

PhD. Professor João Carlos Ferreira, Assistant Professor,
ISCTE-IUL

PhD. Professor Vitor Basto-Fernandes, Assistant Professor,
ISCTE-IUL

September, 2020

iscte

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Energy Consumption Forecasting – A Proposed Framework

Hugo Miguel Nogueira Mendes

A Dissertation presented in partial fulfilment of the requirements of the Degree of **Master in Integrated Business Intelligence Systems**

Supervisors:

PhD. Professor Joao Carlos Ferreira, Assistant Professor,
ISCTE-IUL

PhD. Professor Vitor Basto-Fernandes, Assistant Professor,
ISCTE-IUL

September, 2020

Acknowledgements

I would like to express my sincerest gratitude to professor Dr. João Carlos Ferreira and professor Dr. Vitor Basto-Fernandes for their continuous support and guidance, by making sure I had all the needed resources and motivation to finish this master thesis work.

To my family for being supportive on my academic decisions and believing in me.

To my friends for being truly understandable on the less time we spent together.

To my girlfriend for the unconditional encouragement, support and by believing in me more than I did.

All these people contributed, directly and indirectly, for the conclusion of this master thesis as well as this academic phase and without them, none of this would have been possible.

Resumo

Com o desenvolvimento de países subdesenvolvidos e a digitalização das sociedades, é esperado que o consumo de energia continue a apresentar um crescimento elevado nas próximas décadas. Existindo ainda um grande foco em fontes fósseis para a geração de energia, a implementação de políticas energéticas são cruciais para a mudança gradual para energias renováveis e consequente redução de emissões de CO2. Edifícios são atualmente o sector que mais energia consomem.

De forma a contribuir para uma melhor eficiência no consumo de energia foi proposta uma framework, a aplicar em edifícios ou apartamentos, para possibilitar aos utilizadores ter um conhecimento do seu consumo de energia bem como a previsão desse mesmo consumo.

Diferentes técnicas de análise de dados para séries temporais foram utilizadas para proporcionar informação ao utilizador sobre o seu consumo de energia bem como a validação de caraterísticas importantes dos dados, nomeadamente a verificação da estacionariedade e a existência da sazonalidade, que terão impacto no modelo de previsão.

Para a definição dos modelos preditivos, foi feita uma revisão de literatura sobre modelos utilizados atualmente para previsão do consumo de energia e testados três modelos para os dois tipos de dados, univariados e multivariados. Para os dados univariados os modelos testados foram SARIMA, Holt-Winters e LSTM e para os dados multivariados SARIMA com variáveis exógenas, Support Vector Regression e LSTM. Após a primeira execução de cada modelo, foi feita uma otimização dos modelos para concluir na melhoria dos resultados previstos e na robustez dos modelos para posterior aplicação na framework.

Palavras-Chave: consumo de energia, previsão, framework, análise de dados

Abstract

With the development of underdeveloped countries and the digitization of societies, energy consumption is expected to continue to show high growth in the coming decades. While there is still a strong focus on fossil fuels for energy generation, the implementation of energy policies is crucial to gradually shift to renewable sources and the consequent reduction in CO2 emissions. Buildings are currently the sector that consumes the most energy.

To contribute for a better energy consumption efficiency, it was proposed a framework, to be applied to buildings or households, to allow users to know their energy consumption and the possibility to forecast it.

Different data analysis techniques for time series were used to provide information to the user about their energy consumption as well as to validate important data characteristics, namely stationarity and the existence of seasonality, which can have an impact in the forecasting models.

For the definition of the forecasting models, state of the art was done to identify used models for energy consumption forecasting, and three models were tested for both types of data, univariate and multivariate. For the univariate data, the tested models were SARIMA, Holt-Winters and LSTM as for the multivariate data, SARIMA with exogenous variables, Support Vector Regression and LSTM. After the first execution of each model, hyperparameter tuning was done to conclude on the improvement of the results and the robustness of the models for later application to the framework.

Keywords: energy consumption, forecasting, framework, data analysis

Contents

List of Tables

List of Figures

Acronyms

AR – Autoregressive

ACF – Autocorrelation Function

ADF – Augmented Dickey-Fuller

AIC – Aikaike Information Criteria

ANN – Artificial Neural Networks

BIC – Bayesian Information Criteria

BIM – Building Information Modelling

DECO – Associação Portuguesa para a Defesa do Consumidor

DR – Demand Response

ERSE – Entidade Reguladora dos Serviços Energéticos

GAM – Generalized Additive Model

IoT – Internet of Things

LSTM – Long Short-Term Memory

MA – Moving Average

MAE – Mean Absolute Error

MAPE – Mean Absolute Percentage Error

MSE – Mean Square Error

Mtoe – Millions of Ton Equivalent

PACF – Partial Autocorrelation Function

PV - Photovoltaic

PJ – Peta Joules

RMSE – Root Mean Square Error

RSS – Residual Sum of Squares

SVM – Support Vector Machines

SVR – Support Vector Regression

U.S. – United States of America

U.K. – United Kingdom

Chapter 1 – Introduction

1.1. Overview

According to the International Energy Agency 2018's global energy forecast report (World Energy Outlook), it is expected that energy demand in 2040 will reach 3743 million tons oil equivalent (mtoe), an increase of 27% when compared with 2017. It is also reported that developing countries will have a great impact in electricity consumption where it is expected that Africa has a growth of 140%, followed by the Middle East, Asia Pacific and Central and South America with 96, 84 and 68 per cent, respectively.

With a scenario like this and with the observations by the scientific and geologic communities about the depletion of the ozone layer, the harsh truth of finite sources like oil and gas and the melting of glaciers, forecast a challenging future if actions are not taken at this present moment.

Society is still very dependent on fossil fuels wherein 2018, the total world consumption of oil and natural gas accounted for 4662.1 and 3309.4 Mtoe, respectively [1]. Renewable sources only accounted for 561.3. Without greenhouse policies influencing society to shift to renewable resources [2] and fossil fuels being the cheapest source of primary energy [3], it is expected that the consumption of this sources of energy will again increase in the future. However, the continuous development of new technologies can provide the means for this shift to occur, allowing the production of electricity from renewable sources to become cheaper and thus contributing for a decrease of greenhouse gas emissions.

Although that scenario is still not a reality, it is becoming more feasible throughout the years. According to IRENA (International Renewable Energy Agency), the costs of new technologies to generate energy from wind and solar sources have been declining from 2010 to 2018. Also, since 2010, the global weighted-average of levelized cost of electricity (LCOE) from renewable energy sources such as biomass, geothermal, hydropower, wind projects in nearshore or offshore, has been within the range costs of power generation from fossil fuels, with an interval between 0.49 and 0.172 USD per kWh [4].

The increase of electricity consumption, for countries to continue developing and to maintain their societies, together with the electricity needs of developing countries, will create a problem of an energy shortage, in case the renewable sources issues are not addressed.

The introduction of renewable energy resources provides a significant number of benefits such as the decrease of fossil fuel dependency of countries as well as the decrease of greenhouse gases. However, it is also true that some obstacles appear by its application. The intermittency of energy generation [5] and the high investment required [6] are a few of the barriers that seem to be the most frequent, although others appear depending on the energy source, such as wind farms where it has been shown to have local resistance in countries like Ireland. Despite the country having high winds that would promote a high generation of clean energy, local communities see wind farms as having a negative impact in biodiversity, noise pollution and the decrease of residential property prices [7].

While the scenario of a society that can only generate energy through renewable sources is far from achievable, policies to mitigate the impact of power generated from fossil fuels have been designed and implemented by several countries. In the case of the European Union, the 2020 roadmap defined three pillars of action: 1) reduction by 20% of greenhouse gases at levels registered in 1990, 2) increase energy efficiency by 20 % and increase the contribution of renewable energy technologies by 20% in the final energy consumption [8].

Some sectors, such as the building sector, can have a significant impact on achieving the roadmap goals. In Europe, the building sector accounts for 41% of the overall energy consumption. Households account for almost 27% of total energy consumption, and within this consumption, 71% is used for heating [8].

In order to improve energy efficiency in this sector, some studies have been done to identify key variables that could explain the energy consumption and serve policy-making purposes, change construction processes or develop new technologies that could increase energy efficiency. Some of these variables are ground floor area, building age, window to wall ration, thermal transmittance values to envelop elements, number of occupants, operational hours, [9] environmental temperature [10] and weather [11], to name a few. The identification of the importance of these variables is essential for the correct forecasting of energy consumption. Forecasting energy consumption is a process that can be applied both on the demand side as well as on the production side. By having a good estimate of future energy consumption by different end-users, energy producers can make better predictions of how much energy they need to generate and send do the energy grid in order to meet demand. This information is also helpful for energy producers to allocate resources, minimizing costs and, consequently, reduce energy waste and CO2 emissions, since energy is only being produced to meet the forecasted demand.

Usually, forecasting of energy consumption is also associated with time-series, where a record includes the energy consumed or generated at a specific time interval (hourly, daily, weekly, etc.). Time-series forecasting has been an area of study of more than 25 years and several techniques can be applied for that purpose, such as exponential smoothing, ARIMA models, nonlinear models, long memory models or even ARCH/GARCH models [12]. Each model can help solve different problems regarding time-series forecasting. The robustness and viability of the results depend on factors such as features included in the models and the length of the historical data available.

1.2. Motivation

In [13], the authors identified that the energy sector will be impacted by climate change, being currently one of the biggest contributors to that occurrence. With the increasing awareness of climate change, some works have also been developed to analyse and understand the impact of climate change in our current way of living and how societies interact and manage resources. The same authors analysed several works and identified some impacts that climate change can have in the energy sector: 1) changes in cooling efficiency of thermal and nuclear power generation, 2) changes in seasonal river flows that will impact the potential of generating energy from hydropower, 3) changes in productivity of crops for bio-energy and 4) vulnerability of energy-related infrastructures to extreme events and the rise of sea level. Also, the authors contributed to this area of study by identifying other impacts of climate change for the energy sector, namely changes in space heating and cooling requirements. They concluded that, at a regional level, the changes were more impactful in terms of energy capacity to satisfy additional cooling services, which would result in increases in electricity prices.

The contribution of works in this area is essential for the continuous increase of climate change awareness and help societies to have a proactive attitude by presenting actions while the impacts are still relatively low. The use of sensors to register energy consumption at an hourly basis or a daily basis in conjunction with the application of techniques for energy production and energy demand forecasting has an important role in mitigating climate change. Having a certain degree of information regarding energy that will be consumed and considering information related to external factors, like the speed of wind and temperature, will lead to more efficient resources allocation. It will help energy producers to carefully plan how much energy must be injected into the grid and select which energy sources to use.

The application of machine learning algorithms for time series prediction has been providing excellent results in the past few years, allowing for the planning of energy distribution and generation to be closer to reality.

More importantly, there is a need for consumer awareness when doing simple actions like turning a light on, or increasing and decreasing the temperature of air conditioners depending on the temperature of the room they are in. These actions impact not only the resources needed for that simple action but also the cost the consumer will have by increasing the temperature by three of four ºC when it could be comfortable with only an increase of one or two ºC.

Furthermore, according to the Entidade Reguladora dos Serviços Energéticos’s (ERSE) report on price comparison for domestic consumers, with the information provided by the statistical office of the EU (EUROSTAT), Portugal has an higher electricity price when compared with the average price of the 28 European Union countries where around 49% of that price are taxes, [14]. Additionally, the household expenditure on housing, in Portugal, in 2018 (which considers expenditures with housing, electriticy, water, gas and other fuel sources) was around 17% [15], showing that the weight of energy costs on total income of families is still very high.

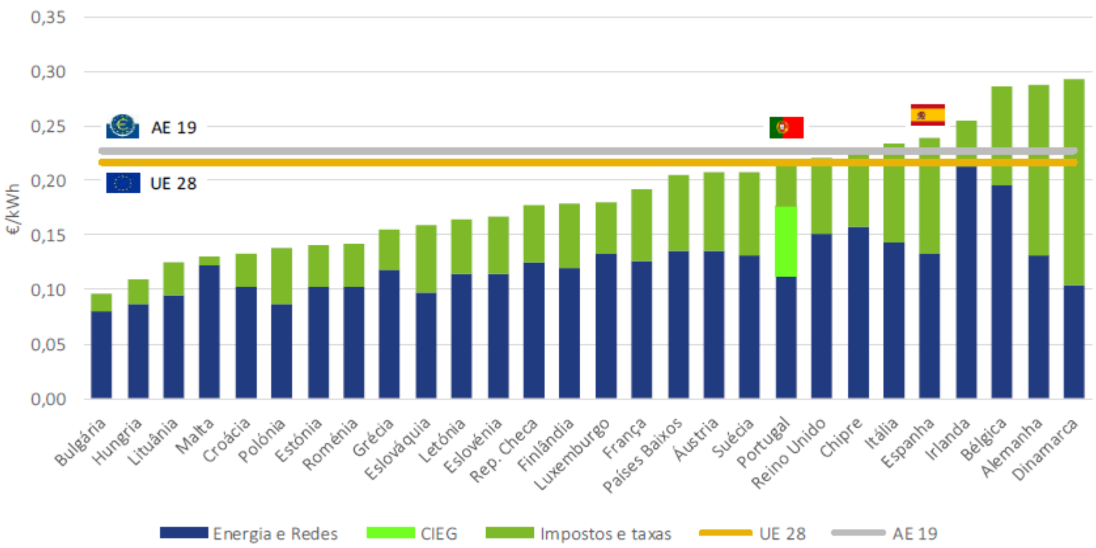

Figure 1 - Electricity price decomposition for domestic consumers [14]

The focus of this work is to contribute to a growing area of study which is justified by the increasing awareness of climate change, finite fossil energy resources, development of developing countries and increase in world population. More specifically, this work is focused on the demand side of energy consumption in buildings and households since the building is identified as having the highest share of total energy consumption, with the proposition of a data analysis and predictive energy consumption framework based on local historical energy consumption data to provide information that could help consumers define and apply policies and strategies to decrease their kWh consumption, respective costs and promote to a more efficient and reliable energy grid.

1.3. Objectives

The objective of this work is to answer the investigation question: “Is it possible to generate approximate local predictions in real-time to provide to the end-user information about the impact on their energy consumption behavior?”. This question is answered with the proposition of a framework with two objectives in mind:

- Provide information on historical energy consumption through data analysis techniques and visualizations;
- Provide approximate hourly energy consumption forecasts based on local historical data with additional information regarding the respective energy costs.

Personalized forecasted energy consumption is generated by using an approximation based on local historical data collected from the building or household. The viability and robustness of the predictive model are better with higher volumes of historical data since trending and seasonality in time-series forecasting has a significant impact on the accuracy of the forecasting models.

The framework provides end-users with information on their energy consumption behavior, based on historical consumption patterns as well as approximate forecasted hourly values. With the possibility of using external data, such as temperature or humidity, the framework will use a predictive model for univariate and multivariate data. The predictive models are identified and tested in Chapter 5 of this document.

Most cooling and heating systems provide an approximation of the energy it will be consumed by year, based on several variables. For example, the Associação Portuguesa para a Defesa do Consumidor (DECO), besides providing tips on how to choose heating and cooling systems, also provides an online tool that helps identify the best systems for heating and cooling based on the geographical location, climatization period (if only Winter or Winter and Summer), type of building, number of rooms to climatize, respective areas, the level of isolation presented in the house or building, the cost interval the person wants to spend among other variables. The selection of different options provides suggested systems and a comparison between them, with the identification of the cost of acquisition and overall yearly cost of energy consumption (in euros).

The proposed framework tries to move away from this traditional approach by using the historical energy consumption data from each location to provide personalized approximate forecasts based on different energy consumption behaviors.

The conceptualization of the framework, as well as the data analysis and predictive model testing is described and analyzed with the data retrieved from a private kindergarten.

This work is integrated into a local ISCTE project: Social_IoT - University Community Engagement in Technologies for Sustainability: a Social Architecture. The goal is based on this local, personalized predictions influence user behavior about the increase/decrease local temperature, turn on/off the light.

1.4. Methodology

For the identification of the problem, algorithm selection and predictive model definition, a well-known framework was used. The CRISP-DM framework stands for Cross-Industry Standard Process for Data Mining and has been widely accepted throughout all industries regarding data mining projects.

In this framework, six phases are defined where each phase has a predefined sequence and a set of guidelines that help understand what should be done in each phase. Although being sequential, the order of the phases is not strict.

The general schema of the CRISP-DM framework is shown in Figure 2:

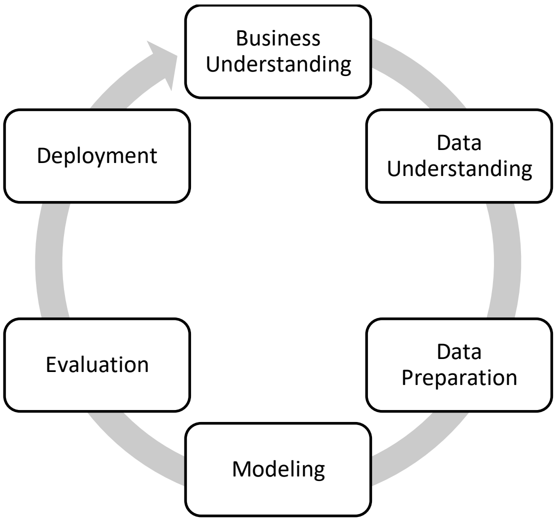

Figure 2 - CRISP-DM Framework [16]

This work started with the Business Understanding stage. In this stage, the focus is to define what we want to achieve by setting objectives and the desired outputs for our problem at hand. Being the objective of this work the definition of a framework to provided information regarding energy consumption and generate approximate local forecasts, initially this stage involved the analysis of the kindergarten data to define the data requirements. The data requirements were related with data types, the granularity of the data (for example, by the hour, by day, etc.), which type of data had to be collected (for example, only kWh data or other) and minimum data periods required for the generation of local forecasts.

For Data Understanding, an exploratory data analysis is done to identify the quality of the data and retrieve general insights from it.

Summing up, the proposed framework and data requirements, were defined based on the data collected from the kindergarten. Chapter 1, 2, and 3 describe both mentioned phases.

In the Data Preparation stage, the focus is to do all necessary data transformation, such as adding new data or data cleaning, necessary to be applied for visualization or to be fed into a predictive model. This stage is described in Chapter 3 and tested in Chapter 4.

At the Modeling stage algorithms are selected and applied to the transformed data, resulting in several interactions and parameter adjustments, until the results provided are satisfactory. This phase is described in Chapter 5.

In the Evaluation stage, the focus is to do the assessment of the data mining results and predictive models and make conclusions on them. In case the results do not respond to the requirements and objectives defined in the Business Understanding stage, there is a need to return to previous stages and redo all the work. This stage is described in more detail in Chapter 4 and 5.

Lastly, in the Deployment stage, the solution is prepared for the go-live, with following monitoring and maintenance. Since this work only focuses on the conceptualization of the framework without the development of software to show the execution from end to end, this stage will not be presented.

Chapter 2 – Literature Review

In this chapter, a literature review was done regarding energy consumption forecasting techniques, energy consumption forecast in buildings and also existing tools that could respond to the research question. The search covered words related to “energy consumption forecasting techniques”, “energy consumption forecast in buildings” and “energy forecasting tools” and was done using the b-on platform as well as Google Scholar.

Doing a search, since 2016, with the words “energy consumption forecasting techniques” on Google Scholar, until March 28th, 2020, 64100 records were obtained. When searching with the words “energy consumption forecast in buildings,” the results included 42600 works. As for the words “energy forecasting tools”, 19 results were obtained.

By removing the year filter and doing the same previous search, the results obtained were 504000, 234000 and 39, respectively. The results show that there are many works regarding energy consumption forecast as well as energy consumption forecast in buildings, which could be explained by the importance energy consumption in buildings have in total energy consumption in a country.

Using the same experiment of analysis at b-on search tool, when searching with the words “energy consumption forecasting techniques” and filtering only by “academic journals”, the results obtained reached 52125 records. When searching for only the words “energy forecasting consumption in buildings”, 25184 results were obtained. As for the words “energy forecasting tools”, the b-on search tool only returned eight results.

Although showing different results, both platforms deliver a high number of works regarding energy consumption forecast and energy consumption forecast in buildings, indicating the importance for researchers continue investigating new ways to forecast energy consumption, while for energy forecasting tools, although there are not many works published, there are many tools in the market for that purpose.

2.1. Energy Consumption Forecasting Techniques

When referring to energy consumption forecast, it is assumed that we are working with time-series forecasting. In time-series analysis, the objective is to extract information by analyzing a sequence of data through chronologically ordered points, allowing for forecasting future values, based on the already registered ones, as well as identifying any past behavior [17]. Data can be divided into univariate and multivariate data. Univariate data is described as only considering one variable for analysis. In other words, the variable itself contains the information needed to be forecasted. Following the same logic, multivariate consider other variables that could be explained and help more accurately forecast the dependent variable. Some predictive models may work better when applying univariate data and have poor performance when applying multivariate data.

The value of energy consumption can be measured, for example, per hour, per day, per week or even per month. By having this type of data, the forecasting of energy consumption values can be done by applying time-series techniques that have been developed and used by several scholars and researchers throughout the years.

“Energy consumption” or “Energy Demand” forecasting has been an area of study by many scholars, researchers and people working in the energy industry. By having information in advance of the energy that will be consumed by individuals at the household level or industry level, helps energy producers and governments to allocate resources better, increase energy efficiency and reduce energy waste. In more detail, energy consumption/demand management is crucial for: 1) planning, 2) energy resource prioritization and optimized utilization, 3) defining policies and strategies for emissions reduction as well as for energy management in the consumer side [18].

In terms of time-intervals, the forecasting of energy consumption can be divided into three categories: 1) short-term, 2) medium-term and 3) long-term. The division into these three categories is based on the information predicted. If the prediction is based on data intervals between one hour and one week, it is considered short-term; between one week and one year, medium-term; and higher than one year, long-term [19].

In reference [12], there has been work done regarding time-series forecasting for at least 25 years. The authors gathered several papers throughout those 25 years and identified a set of techniques and the corresponding variations that have been used for time-series forecasting.

One of the most commonly used and known forecasting technique is the ARIMA model. In [20], the authors applied the ARIMA, the GM (1,1) and a hybrid technique involving the previous models, to forecast China's primary energy consumption. The study covers the rapid increase in energy consumption in China related to its development that could have a global impact on the global energy market. The authors concluded that while giving different results in terms of predictions, all three models were suited to forecast primary energy consumption in China.

In [21], the authors applied the ARIMA model to forecast energy consumption and GHG emissions from a pig iron manufacturing organization in India, since the managers wanted to know the trend of those metrics in order to better apply environmental policies. For the energy consumption forecast, the most suited ARIMA model was developed with the parameters (1,0,0) x (0,0,1), while the most suited ARIMA model for GHG emissions was developed with the parameters (0,1,4) x (0,1,1). Further analysis of the information provided showed a decreasing trend of both energy consumption and GHG emissions.

In [22], the authors wanted to forecast energy consumption using two predictive models: ARIMA and an Auto-Regressive Neural Network (NAR) model. The authors concluded that both models presented results that allowed them to consider both as suited for predicting energy consumption.

Another technique that can be used for time-series forecasting is the exponential smoothing Holt-Winters model, which is an extension of the Holt's linear equation model, adding the capacity to capture the seasonality of the data. The model helps to forecast short, medium and long-term values and has two versions for its application: additive and multiplicative. The multiplicative model does not provide good results when null or negative values are present in the data [23]. The characteristics of this model make it suitable to be applied for time-series forecasting, especially for energy consumption forecasting.

The application of this model was used in [23] to compare two linear models: ARIMA and Holt-Winters for forecasting Primary Energy Consumption Total data in the U.S. The comparison was based on the evaluation metrics MAE (Mean Absolute Error), RSS (Residual Sum of Squares), MSE (Mean Square Error) and RSME (Mean Square Error) and with data ranging from January 1973 to December 2016. The authors concluded that the Holt-Winters model, with the additive component, provided more accurate forecasts when compared to the Holt-Winters model with multiplicative component and the ARIMA model, with MSE values of 258350.1, 262260.4 and 723502.2, respectively.

Similar to the previous work, in [24], the authors also compared the models ARIMA and Holt-Winters to forecast total and component-wise electricity consumption in Pakistan. The data ranged from 1980 to 2011 and covered the consumption of electricity in household, commercial, industrial and agriculture sectors, to name a few. With the application of these two models, the authors concluded that the Holt-Winters model provided the best results when compared with the ARIMA model and identified that forecasted increased of electricity consumption would generate a possible problem in the future with the low growth rate in electricity supply in that country.

When dealing with multivariate data, the ARIMA models is not the most suitable for forecasting. To have into account other variables, other modifications of the ARIMA can be applied. In [25], the authors compared four methods for electricity generation forecasting of grid-connected Photovoltaic (PV) plants in Greece, in forecasting the day-ahead and intraday values, being these four models identified as SARIMA, SARIMAX, modified SARIMA and ANN. The SARIMA model, a variation of ARIMA, is classified as being a linear approach for time-series forecasting and with a characteristic for being able to remove seasonality from data through seasonal differencing [26]. When including the X to SARIMA, we get SARIMAX, a seasonal ARIMA model that include exogenous variables (X), thus allowing the ARIMA model to be able to be used with multivariate data. The conclusion of this work was that for the day-ahead, the SARIMAX, ANN and modified SARIMA provided the best performance results, with a normalized RMSE of 10.93, 11.42 and 11.12, respectively. As for the intraday, the SARIMA (univariate data) performed better than the SARIMAX (multivariate data), with a normalized RMSE of 8.12 and 9.11, respectively.

In [27], the authors applied different predictive models to forecast daily and monthly snow water equivalent in Ontario, Canada. For daily forecasts, the predictive models used were ARMA and ARMAX, as for monthly forecast values, SARIMA and SARIMAX were applied. For a daily forecast, the authors concluded that the ARMAX model with time trend component (TT-ARMAX) provided the best results, with an RMSE of 2.52. As for the monthly forecasts, it was observed that by adding exogenous variables, the SARIMA model improved its performance.

Also, in [28], the authors applied forecasting methods for daily natural gas consumption on a regional basis in Turkey. These models were SARIMAX, ANN-MLP and ANN-RBF. The authors concluded that SARIMAX had fairly robust results for short-term local forecasting of natural gas consumption in Sakarya providence in Turkey, providing an RMSE value of 0.122, while the ANN models presented an RMSE of 0.44 (ANN-MLP) and 0.50 (ANN-RBF).

Non-linear models have also shown good result for the forecasting of energy consumption.

Variations of the Support Vector Machines have also been applied in different studies of energy forecasting, more specifically Support Vector Regression (SVR). The SVR can be applied not only to univariate but also with multivariate data [29][30][31]. In reference [30], SVR eliminates restrictions observed in SVM techniques such as only be applied to classification with the output variables only taking binary values. The SVR allows for the usage of non-linear functions. In the same work, the authors applied two techniques, Neural Networks and a seasonal SVR to forecast electricity consumption of Turkey. Using the MAPE as the evaluation metric, the ANN and seasonal SVR presented a result of 13.8% and 11% in the training dataset, respectively. As for the test dataset, considering two years of data (2010 to 2011), the results were 3.9% and 3.3%. The conclusion was that the seasonal SVR model outperformed the ANN model and analyzing by analyzing the results it provided empirical evidence that electricity consumption in Turkey could be forecasted using an SVR model.

SVR models can also be applied to forecast power generation. In [31], the authors applied a model-based SVR and an ANN model to forecast the power generation of three different PV stations using historical PV power output and meteorological data. The evaluation metrics used were normalized RMSE, MAE and mean bias error (MBE) and both models were compared to a persistence model used as a benchmark for forecasting PV power. By using three different months from 2016 (January, May and September) and averaging the results, the author concluded that the SVR model outperformed the ANN model and the benchmark model, with an average normalized RMSE, average MAE and average MBE of 3.08, 34.57 and 11.34, respectively when compared to the results of 2.86, 48.83 and 13.58 from ANN model and 13.23, 100.93 and 46.88 from the benchmark model, thus being an accurate model for PV power output forecasting.

Artificial Neural Networks, a non-linear model, was used to forecast electricity consumption in Thailand, in [32]. The author compared the results obtained from applying an ARIMA model, an ANN model and a Multiple Linear Regression (MLR) to a dataset that contained the yearly historical data from 1986 to 2010 of population, GDP, SET index, Export (in a million baht) and Electricity Consumption (in GWh). The author concluded that, by analyzing the MAPE, the ANN model had the best performance when compared with the ARIMA and MLR, although in terms of errors the test did not show a difference between models. The author also concluded that the accuracy of the ANN model might be jeopardized by overfitting because of the limited number of available training tests.

In [33], the authors used two models of ANN to forecast energy consumption in South Africa's industrial sector by using yearly data from 1993 to 2000, which contained the GDP variable and the total energy consumption, measured in PJ. The ANN models applied were the Multilayer Perceptron (MPL) and the Radial Basis Function (RBF). Using the evaluation metrics of R^2 and MAPE, the authors concluded that the RBF provided more accurate results and it could be applied to forecast South Africa's industrial sector energy consumption and help formulate a viable energy strategy.

In [34], the authors applied ANN methods in order to identify the influence of National Income, Population, GDP and Consumer Price Index on electricity consumption in Taiwan and developed a forecasting model. The author concluded that Population and National Income had the most significant effect on electricity consumption and that ANN showed better results in terms of the RSME, MAE and MAPE metrics, which can be more suitable to develop a forecasting model for Taiwan's electricity consumption.

2.2. Energy Consumption Forecast in buildings

"Energy consumption" or "Energy Demand" forecasting has been an area of study by many scholars, researchers and people working in the energy industry. By having information in advance of the energy that will be consumed by individuals at a household level or by an industry level, helps energy producers and governments to allocate resources better, increase energy efficiency and reduce energy waste. Energy consumption/demand management is crucial for: 1) planning, 2) energy resource prioritization and optimized utilization, 3) defining policies and strategies for emissions reduction as well as for energy management in the consumer side [18].

The research on time series forecasting has been registered for over 25 years by the International Institute of Forecasters that helped develop forecasting as a multidisciplinary field of research and channeling the knowledge on predicting in order to help society. In [12], the authors analyzed over 25 years of papers, key papers and books related to time-series forecasting and developed a review work about the models that can be applied for time-series forecasting. In this works Arima models, exponential smoothing, non-linear models, long memory models and ARCH/GARCH models were identified as being tested in time-series datasets with results that would promote them as being adequate for the analysis at hand. Additionally, in this same work, some forecasting evaluation and accuracy metrics were also identified. Metrics such as MSE, RMSE and MAE were classified as being some of the most commonly used forecast accuracy measures in works regarding time-series forecasting.

In [35], several forecasting methods were analyzed and categorized by: Complexity, Easy to use, Running speed, Inputs needed and Accuracy. While statistical models were categorized has been fairly complex, easy to use and with reasonably high speed, they lack in terms of accuracy where it was classified as being fair. On the other hand, ANN and SVM while being categorized as being complex, not easy to use, it gives fairly high and high values of accuracy, respectively.

When validating the real applicability of the forecasting models, several papers can read and analyzed with interesting results and conclusions. In [26], the authors proposed a model for combining SARIMA and ANN model to forecast the annual energy cost budget through the analysis of energy consumption registered in several educational facilities in South Korea. This proposed model was also compared with the conventional models such as the SARIMA model and the ANN model. The data was comprised of 7 years of electricity consumption (from 2005 to 2011) from 787 educational facilities. It was concluded that the proposed hybrid model showed a MAPE mean absolute percentage error between 0.11 and 0.23% when compared with the 1.23 and 1.84% from the conventional SARIMA model. In both cases, it can be derived that the SARIMA model, being directly applied for energy consumption forecasting or by being integrated into a hybrid model with different predictive models, is a reliable predictive model.

In [36], the authors used the SARIMA model to forecast short-term electrical load data. The results obtained proved that the SARIMA model could be applied for energy consumption forecast although with better results with monthly load data instead of hourly load data.

In [37], the authors also used the SARIMA model to forecast an office building elevator and compared the results with an ANN model and a GAM model. The predictions were made in a daily, hourly and 15-minute basis, and the models were able to identify the seasonal patterns in the data. When comparing the three models, all showed good results which were acceptable for energy consumption forecasting. In the end, the SARIMA model performed better than the ANN and GAM despite being a simpler model.

In [38], the authors applied 3 predictive models to forecast electricity consumption for Pakistan. They used both linear and non-linear modelling techniques which included ARIMA, SARIMA and ARCH/GARCH models. The data analyzed was registered over a period of 11 years, from January 1990 to December 2011. The results from all models let them conclude that the ARIMA provided the best value for the MAPE evaluation metric followed very closely by the SARIMA model tested, thus identifying that these models were reliable to be used for electricity consumption forecasting.

Also, in [39], the authors applied two predictive models, one linear and one non-linear model, to forecast energy consumption using data from the Federal University of Mato Grosso, in Brazil. The models selected by the authors were SARIMA and autoregressive neural network, NARNET. The authors concluded that both models had the potential to solve the problem of identifying the energy required to satisfy the needs of the Federal University of Mato Grosso to function appropriately with the lowest energy cost required.

In [40], the authors applied an SVR model to forecast short-term electrical load and calculate demand response baseline in several office buildings, by considering temperature as an input variable. The new SVR model proposed would use a two hours temperature period before the Demand Respond (DR) events occur in order to improve the forecasting accuracy. The model was compared with the other seven methods for DR where it shows better results than the other models, by achieving a MAE of 1.57% in their four office buildings used for testing.

In [41], the authors used the SVR model as well for building energy consumption forecast by using hourly energy consumption data in two different buildings for one month. The first three weeks were used as training test as the last week as a test.

Additionally, the authors used traditional time-series models (ARIMA) to compare the performance of the SVR model. By using MSE as the evaluation metrics, the results showed that SVR performed better than the ARIMA model with 0.01 and 0.07 for building 1 and 2, respectively opposed with the 0.02 and 0.09 from the ARIMA model, concluding that the SVR was a suitable model to forecast energy consumption.

In [42], the authors compared a random forest predictive model with an ANN predictive model on data regarding the electric consumption of a heating, ventilation and air conditioning system (HVAC) from a hotel building. This data was gathered from its building management system. Additionally, the data also had records of outdoor air temperature, dew point temperature, wind speed and relative humidity, as well as more hotel-related data, such as room occupancy. The authors were able to conclude that the ANN model provided best evaluation metric values when compared to the random forest model, with an RMSE of 4.97 and 6.10, respectively, thus helping managers to act with more reliable information regarding the hotel’s energy consumption.

In [43], the authors analyzed the results of applying an ANN with Bayesian regularization algorithm for forecasting sub-hourly electricity usage in a commercial building. By adding key variables, such as day type indicator, time-of-day, HVAC set temperature schedules, outdoor air dry-bulb temperature and outdoor humidity into the model and by using adaptive training methods, the authors concluded that the forecasting model was able to predict electricity consumption within a 15 min time frame, for example.

Also, in [44] an ANN model were proposed to forecast seasonal, hourly electricity consumption in three areas of a university campus in Japan. To develop the model, the authors identified vital variables such as day of the week, the hour of the day, hourly dry-bulb temperature, hourly relative humidity, hourly global irradiance and the previous hourly electricity consumption. While considering the feed-forward ANN in conjunction with the Levenberg-Marquardt back-propagation algorithms, the results obtained of R^2 and RSME ranged between 0.95-0.96 and 0,99, respectively. Although showing excellent results, the authors considered that by adding more variables to the model could improve its accuracy, such as information of indoor human activity.

All these works contributed to the recognition that linear and non-linear models can be applied to forecast time-series, especially energy consumption through time.

Depending on the model used and the characteristics of the data, the results can be better or worse.

To sum up, these works show that the models for forecasting energy consumption are robust and viable.

For this work, predictive models for univariate and multivariate data have been tested, based on the literature review done in this chapter. With the proposition of a framework for energy consumption analysis and forecasting, it was considered to be more effective to use some traditional and non-linear forecasting models, which have been scientifically proven to be suitable for time series forecasting, in order to identify which should be the most suitable one to be applied to different cases while providing good results.

2.3. Energy Forecasting Tools

In terms of energy forecasting tools, there are very few works that provide and show results on the application of a forecasting tool for energy consumption. In the work of [45], the authors had two goals. The first one was to develop a methodology for hourly energy forecasting by using diverse machine learning techniques, such as Support Vector Machines, Trees and Neural Networks. The second goal was to propose an intelligent system, named VIMEON, to enable three-dimensional data visualization and forecasting of energy consumption in real-time. The feasibility of the proposition was tested using data from different buildings, associated with the University of Granada, situated in different cities in Spain. The authors concluded that the proposed model provided accurate results of energy consumption forecasting as well as a simple and very interpretable system for data visualization. At a macro level, the proposition is very similar to the proposition we are doing in this work, however, it differentiates from one another in terms of machine learning techniques applied as well as technologies used.

In [46], the author analyzed several tools and services focused on load and renewable power, categorizing them and providing general information of principals of time-series and energy consumption forecasting. The authors concluded that most of the tools and services provided have to be purchased, come from private companies and focused more on the load forecasting side instead of the volatile renewable power forecasting. Due to the recent publishing date of this work, it was considered that this information was very important for the proposed framework identified in this work since it helps for comparison, by identifying several tools for energy forecasting, the horizon level of the forecasting and some of the machine learning techniques used. The authors also report that the information gathered was based on the publicly available information displayed online, without having contacted someone from that specific company. Furthermore, it is not clear if all the tools and services identified are for companies and/or for household habitants or casual users.

To complement with this information, some tools were selected for further understanding of how they work, source types for data gathering and machine learning forecasting techniques used.

The first software identified that can be used to forecast energy consumption is provided by the GMDH company, the GMDH Shell. According to the company's description, this software applies different methods, such as linear, polynomial, Gaussian or neural networks, to build the predictive model. It provides different templates for the time intervals pretended for the forecasting as well as the option for the user to indicate the number of observations he wants to forecast. The results are provided in an interface that comprises graphs and tables, with the real values, the forecasted values, the lower and upper confidence intervals and the accuracy values for the model.

SAS also provides software, named SAS Energy Forecasting, for load forecasting and energy trading with a list of features that promotes all levels of decision making to operate more efficiently and effectively. According to its user guide, the forecasting process is composed of two stages: 1) a diagnose process and 2) selecting forecasted period, between short/very short-term and medium/long-term. In the diagnose process, the first step is to create a naïve model that will be the benchmark for the next stages, where the process then adds other features, like recency effect or weekend effect, calculating the MAPE for each execution for later comparison. When the best model is identified, its residuals are used to create other models, such as, and to name a few, neural networks, exponential smoothing and ARIMA with exogenous input. This software is primarily intended for utility companies, thus justifying some of the observed complexities for using it.

Another company that provides software for electricity demand forecasting AlexaSoft. The company provides to types of forecasting: energy price forecasting and demand and renewable energy forecasting. Their product utilizes AleaModel as a unique type of forecasting model and specified for energy forecasting, by combining the best of several forecasting techniques such as SARIMA, Neural Networks and regression. The focus of their solution is for large, electro intensive consumers, utilities, marketers, transmission system operators, traders, investment funds and banks.

MextrixND is a software provided by Itron and described as the "*Itron's industry-leading forecasting engine allows rapid development of accurate forecasts".* This software has been improved and refined for more than ten years and is user-friendly by applying a drag-and-drop methodology. As for data sources, the software can incorporate data from excel spreadsheets and databases, such as Oracle or SQL Server. In terms of machine learning techniques this software uses Exponential Smoothing, Regression, ARIMA and Neural Networks to forecast short to long-term information and is mostly used within utility companies, ISOs, municipals, cooperatives and other energy service provides.

Table 1 aggregates the previously mentioned companies and software as well as some of the identified predictive models each utilizes.

Overall and considering only a small sample of available energy consumption forecasting softwares, all of the previous are more focused on companies and / or energy producers, having a level of understanding somewhat complex for the average household person or families. The proposed framework was conceptualized to be able to be used by any user despite having little to no knowledge on modelling and forecasting techniques.

Table 1 - Companies and respective software for energy / load forecasting

Company	Software Name	Predictive Models	Information Source
GMDH	GMDH Shell	• Linear Models • Polynomial Models • Gaussian Models • Neural Networks	https://gmdhsoftware.com/predictive-analytics-software
SAS	SAS Energy Forecasting	• Arima • Exponential Smoothing • Winters Method – Additive • Winters Method - Multiplicative	http://documentation.sas.com
AlexaSoft	AlexaModel	• SARIMA • Neural Network • Regression	https://aleasoft.com/aleamodel/
Itron	MextrixND	• Exponential Smoothing • ARIMA • Neural Network • Regression	https://www.itron.com/na/solutions/product-catalog/energy-forecasting-group

Chapter 3 – Framework Design

In this Chapter, the framework for energy consumption is proposed, with the description of the requirements necessary for its correct applicability and functionality. The framework aims to answer the investigation question defined in section 1.3.

The proposed framework described in the section provides an end-to-end process, meaning that the framework will encompass data extraction, data treatment, data analysis, predictive models, and data visualization. These stages are related with the Data Understanding, Data Preparation, Modeling and Evaluation stages represented in the CRISP-DM model and were conceptualized having in mind specific tools, such as Java for the development of the software that will provide data visualization as well to enable all the action to be done through Python, MySQL for data storing and Python for the data treatment, data analysis and forecasting. In the long-term, in case data storing becomes an issue due to the increase of data, the migration to an Azure database is a possibility since it is a cloud resource with high scalability, computational power and security.

The definition of the framework was initially thought by analyzing the available energy consumption data that was also posteriorly used for testing the applicability of the data analysis component and predictive models of the framework. The data collected belongs to a private kindergarten named “Pequenos Sorrisos, Lda.”, located in Amadora, Lisbon, Portugal (Latitude 38.7664 and Longitude -9.2388) with 50 students. The building has 12 rooms, with dimensions of 16 to 40 m^2 that provides activities for kids of ages between 3 months old to 5 years. In 2016, the kindergarten implemented a sustainability project in order to be more energy-efficient, with the upgrade of the lighting system to LEDs as well as the instalment of a real-time energy monitoring platform available on the market. Additionally, the building also has electric blinds, that when used, will have an impact on the kWh consumption.

Figure 3 - Private kindergarten visual identification

This building was initially a case study for a master thesis elaborated by a student of ISCTE-IUL with the focus on the development of Smart Environment Monitoring using LoRa [47] having the kWh data been provided by that same student. The author concludes in his work that the application of a monitoring system helped the kindergarten owners to identify several situations that had to be addressed in order to decrease energy consumption. Some examples of these situations were circuit breakers of air conditioning units that had a persistent waste of 30 watts even if all units were off, mosquito nets in the kitchen had a consumption of 60 watts or computer monitors with a consumption of 5 watts even when they were off. The author concluded that with his work, some anomalies where identified, is related not only to the building (windows, lighting, etc.) but also to the appliance usage during the night, and with the application of energy-saving policies and specific action for the identified anomalies, in three years there was an energy-saving around 20% to 25%.

The previously mentioned work also did done some exploratory data analysis on the same dataset by using the monitoring system developed, however in this chapter exploratory data analysis techniques were also applied since the objective is to simulate the results the application of the framework will provide to the user. There was no comparison between the data analysis from this work with the results presented in [47]. This data analysis is also used to identify correlations between the variables, that are identified in this chapter, in order to help the user to see if they should include them in the predictive model or not.

The data was collected from various .csv files, with each file representing a month, and having 26.715 rows total with some duplicate values (since each dataset was saved for each month, the last day of each file was also the first day of the next month at hour 00:00:00). A quick analysis on the energy consumption registered showed some hours with missing values (the month of April 2019 did not show kWh consumption from 2019-04-06 18:00:00 until the end of the month and from the beginning of September 2016 until 2016-09-13 18:00:00), as well as some timestamp showing 0 kWh consumption. Assuming that most appliances or types of equipment remained connected even during non-working hours (for example, refrigerators, microwaves or telephones), the kWh consumption should have a value above zero. While the missing values could be explained by errors in the connection between sensors and the data storage, others were more difficult to explain due to the lack of information, such as 0 kWh consumption records. For this work, it was assumed that we could be looking at poor data quality, thus some actions were taken to improve its quality, namely the application of the moving average method to fill the missing data. To fill the missing values, the moving average method was the most suitable option since it uses past values to predict future ones [17] and it was applied with a window of 5, meaning that the specific value of a record in a specific date and time was filled with the mean value generated from itself and the previous two records.

Additionally, it was analyzed that some points in time had very high values that could mean outliers. Generally, outlier identification can be difficult to identify correctly since it depends on the data at hand. In this case, since energy consumption can be unpredictable and depends on the actions people have on the day-to-day activities or by some anomalies that appliances must-have during its operational lifecycle, all data points were considered for the data analysis.

The analysis of this data and respective quality helped for the definition of the framework, with two main objectives conceptualized: 1) apply data mining techniques for data analysis to provide information to the consumer about their energy consumption behavior and 2) apply forecasting techniques, based on the historical data collected and external variables such as temperature or day of the month, to provide information regarding future energy consumption and related costs. These objectives will be reflected as two features on a software to be installed in the user's computer and will be selected by the users depending on their necessities.

Since time-series forecasting needs several data points recorded through time in order to correctly identify patterns and accurate forecasting, this option for consumption forecasting has a rule to only be executed if the there is, at least one year of recorded kWh consumption data.

The framework will be very similar to the already existing products in the market, some of them stated in section 2.3, with some differences in regard to predictive models utilized and added information provided to the user, such as energy pricing.

Since this framework was thought to be used by energy users, it might bring more value if this framework is implemented by a utility company, incorporated as an additional service in the new or existing contract, instead of an independent service by a non-utility company. Furthermore, being delivered by the respective utility company, the price to add to the forecasting values will be more accurate than identifying an average price, since each energy provider would already have pricing information about their clients.

During the explanation of the framework in the following sections, the focus will not be on the development or the in-depth analysis on data extraction systems and techniques but rather on how the process would work in terms of the data treatment, data mining and forecasting techniques. Conceptually, the framework would be defined as presented in Figure 3.

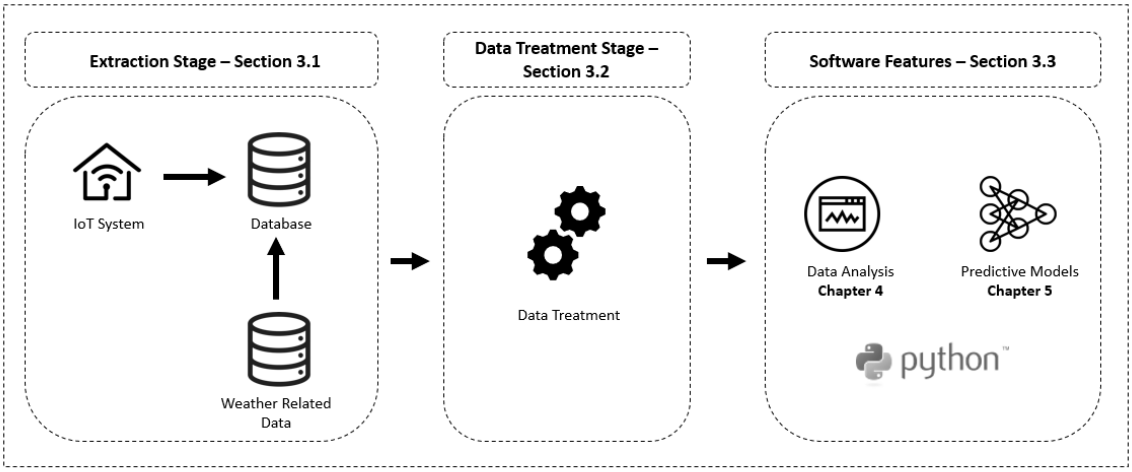

Figure 4 - Proposed framework model

3.1. Extraction stage

Briefly speaking, the extraction of data into the database comprises of energy consumption in kWh and weather data, namely temperature in Celsius, humidity in percentage and precipitation in mm. Energy consumption data, in kWh, is collected using an IoT (Internet of Things) system that registers the energy consumed at an hourly basis, by using a clamp meter in each phase of the electrical panel. Each clamp has a transmitter that sends information to a receiver that is connected to the internet which enables the storage of information at an hourly basis and electric consumption values, in kWh, for posterior data mining and forecasting, in a MySQL database or Azure database, depending on the storage and computational resources necessary. For this version of the framework, the energy consumption data represents the overall kWh consumption of the household or building without specifying the consumption by room, power plug or appliance. However, having this information would help improve the framework since it could give more accurate values do identify consumption patterns in specific parts of the household or building. The data collected must present the timestamp, and the kWh consumed, and as it enters the database, it is divided into two columns: 1) Timestamp and 2) kWh.

For weather data, the extraction process is based on data provided by the website "wunderground.com" since it delivers hourly information which can be linked with the hourly information already stored in the database, based on the closest region available. Since the available APIs identified only provide information regarding future weather and temperature data, the historical weather data is being collected using web scraping techniques from the "wunderground.com" website.

The extraction process has an hourly periodicity in order to have the most recent information available when executing the forecasting models.

As for the pricing costs, energy producers have different pricing contracts that depend on the contracted power, hourly consumed option, payment methods and if it includes other energy services, such as Gas. Considering all these variables, the pricing value is manually inserted into the database based on the average price the client has contracted with the energy company. Since the pricing values are adjusted each year by ERSE, the price will have to be adjusted as well from year to year. In case there is a change in energy provider or energy contract, the pricing will also have to be affected. These changes are not automatic and need manual interaction for the respective update, thus needing user interaction from time to time.

Since the extraction process is crucial for the accurate generation of historical data visualization and forecasting, four conditions during the extraction process must be met: 1) have data regarding kWh consumption at an hourly basis, 2) Have data regarding temperature, humidity and precipitation at an hourly basis, 3) have the average energy pricing under the contract the building or household has with the energy producer and 4) have at least one year of data collected for the correct application of the predictive model.

After storing all necessary data, its representation in the database is as shown in Table 2 and Table 3.

Table 2 - Database structure with an example of a record with data

Timestamp	kWh	Temperatura (Cº)	Humidade (%)	Precipitação (mm)
01-05-2020 10:00:00	0.20	18	60	30

Table 3 - Configuration table with an example of a recorded pricing value

ID	Begin_Date	End_Date	Unit_Price
1	01-01-2019	31-12-2019	0.15

3.2. Data treatment stage

The data treatment stage is executed for both options provided in the software. Python libraries are used not only for the collection of information from the database but also to apply all necessary techniques to the data to be used for later data analysis and forecasting. The *mysql.connector* package enables the connection between Python and the MySQL database by indicating the name of the server, the database name and the query we want to use to extract the data. In case the database is an Azure SQL Database, the library to be used is *pyodbc*.

Considering the importance of historical data for time-series forecasting, in case there are missing data originated, for example, by an anomaly in IoT system, the framework has a control rule to apply the moving average method. As previously stated, the moving average method was considered the most suitable option to fill the missing values for this framework. Other methods, such as forward fill, backward fill or interpolation could also be used for missing data and might have different impacts on the predictive models depending on the data at hand. As for outlier treatment, it is important to note that each dataset will have its own specificities, so the definition of an outlier treatment method that fits all is difficult to define. Time-series outliers can correspond to points in time that occur in seasonal periods, thus being difficult, without further analysis, to understand what action should be taken. Being so, and as stated in the beginning of this chapter, for this framework no actions are defined for outliers, as it can provide important information in identifying appliances anomalies or specific dates where the energy consumption has an abnormal higher value.

Furthermore, the availability of timestamp information also permits the creation of more columns that are used for data visualization as well as features for the predictive models. The new columns created are in reference to the day of the week, if it is a working day, year, month, day of month and hour. At the end of the treatment process, the data will be inserted into the database for further use by the next processes. The columns and respective data types that are used for the data analysis and visualization as well as to be fed into the predictive models are as presented in Table 4.

Table 4 - Columns and respective data types after data treatment process

Column	**Datatype**
Timestamp	timestamp
Temperatura (Cº)	int
Humidade (%)	int
Precipitação (mm)	float
Day of week	int
Working day	boolean
Day of month	int
Month	int
Year	int

3.3. Software features

In terms of software, it is developed with two features in mind:

- Data Analysis;
- Predictive Models.

The first feature presents information and visualizations regarding local historical data. The second feature performs hourly local forecasting kWh consumption values. The "Predictive Models" option will only advance to the next stage if there is, at least, one year of consumption data. The one-year data limit was defined since it shows all the seasonality and trend on energy consumption that occurs throughout the year, which includes all four seasons. However, it is important to be aware that specific events in time that might occur in one year and not on the next can make the model lose some of its robustness since it might try to predict that same event in the next year.

3.3.1. Data Analysis

Describing with more detail, by selecting the first feature, the process applies data analysis techniques that are used to display a series of information to show to the user, namely:

- a line graph showing kWh consumption through time (up to three years);
- descriptive information regarding energy consumption behavior, such as mean kWh consumption, top five max values recorded, and top five min values recorded;
- total kWh consumption by the hour, day of the month, month and year;
- cross-correlation between kWh consumption and other time series variables namely, Temperature, Precipitation and Humidity;
- Pearson correlation between kWh consumption and the categorical and numerical variables, namely hour, day of the month, month, year, day of the week and working day.

This information can also give users knowledge on possible increases in consumption on a specific day of the month, for example, that could help identify anomalies or events that need to be addressed.

The correlation between all variables is presented to help the user understand how the different variables will influence the predictive results. As stated before, two types of correlation are used in the framework. Since it is important to analyze the relationship between variables in order to create a robust predictive model, the method to identify those relationships also needs to be correctly applied. The cross-correlation method is used to identify the relationship between different time series. This method has been applied in [48] and [49].

For the other variables, the Pearson Correlation Coefficient is used by showing how much the strength of the linear relationship is between two variables, ranging from -1 to 1. Values close to 0 indicate that the variables do not have any linear relationship, where values close to -1 or 1 indicate a strong linear relationship between them, being negatively or positively, respectively [50].

3.3.2. Predictive Models

The second feature utilizes processed data to feed into the predictive model and generate local forecasts.

By selecting this feature, first, all variables are displayed in a new window. The user can either select all external data (temperature, humidity, precipitation) and other variables such as day, day of week and month, or only the kWh consumption. In this window, there is a text box indicating that selecting variables that do not show a strong correlation might affect the accuracy of the forecasted values.

After selecting the variables to be included in the model, the user can then identify the number of data points they want to forecast (each point is defined as one hour, so for example, if the user wants to predict the value for the next 24 hours, it has to indicate that they want to forecast 24 data points). Depending on the variables selected, the models may have to change as well, since univariate data works better with certain predictive models while multivariate data perform better with other models. Having this into account, two models are selected to be included in the framework, one for univariate data and one for multivariate data (when considering external variables such as temperature, humidity, or day of the month, for example). Information regarding the predictive models used for the framework are described in Chapter 5 of this work.

With the variables chosen and the data points to be forecasted selected, the process starts by analyzing the stationarity of the data or, in order words, if the data has a positive or negative trend through time. In case the data is not stationary, it will be necessary to apply a differencing method, which is commonly used in time-series forecasting. To be able to identify this condition, the parametric test Augmented Dickey-Fuller test is defined using Python, where it generates a *t-statistic* and a *p-value* that will be used to conclude if the null hypothesis should be rejected or not. In case the *p-value* has a higher value than 0.05, then the data is not stationary, and the differencing must be applied. In general terms, the differencing is the action to values of a time-series into changes of values of a time-series. More specifically, the differenced time-series at value *t* is the result of time *t* minus the values of time *t*-1 [17]. After the application of the differencing in the data, a new stationary test is executed to reconfirm if the value of the *p-value*. The differencing should continue until the data is stationary.

Considering the stationary of the data, the next step is to insert the data into the predictive model to forecast the data points selected by the user. After calculating the forecasted values, the information is provided to the user, through the interface, in a line graph. Figure 5 provides an example on the graph that is presented to the user. Additionally, a table showing the forecasted values and the total costs is presented to the user to keep a track on the calculated values. The information available on the table can be downloaded for posterior analysis, if necessary.

With the forecasted values being approximations on the kWh consumption that will occur in the future, the associated total costs will also be approximations, being above or below the real cost. Table 5 gives an example of the table that is presented to the user with the information regarding the forecasted values and associated costs. This table is also loaded into the database for historical purposes.

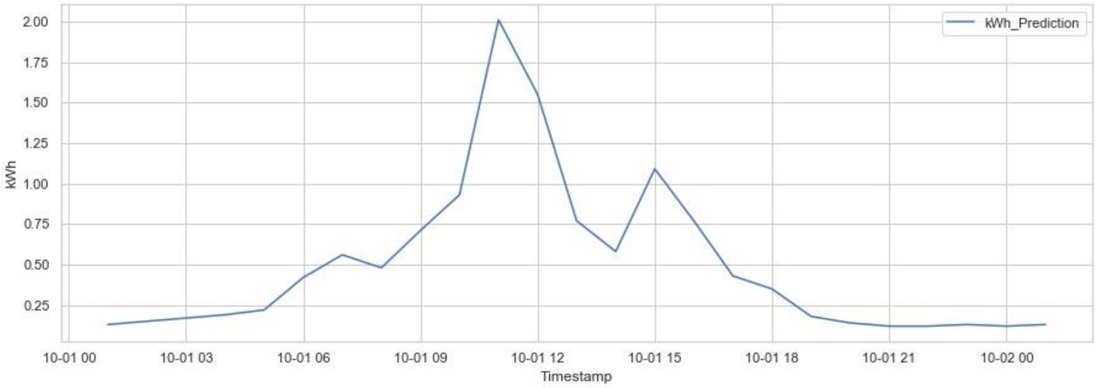

Figure 5 - Graph line presenting forecasted kWh consumption by hour

Table 5 - Table structure of a forecasted value with the associated cost

Date	kWh_Forecasted	Unit_Price	Total_Cost
30-06-2020 15:00:00	0.21	0.15	0.04

Chapter 4 – Exploratory Data Analysis

The objective of this chapter is to simulate some stages of the framework, specifically the data treatment and data analysis stages, and conclude on its applicability of the framework for the objectives at hand based on data collected from a kindergarten. Due to the source type of the data (.csv), some actions are performed in this analysis that will not be used in the framework, namely the removal of duplicate rows found in each dataset. The action taken for this situation is described later in the next section.

4.1. Data Treatment

Specifically for this dataset, in order to prepare the data for the data analysis and the predictive models some additional actions were taken, namely: 1) to export the values from April 2018 to April 2019 only where the data was missing from April 2019, 2) remove from the dataset the values from 2016-09-01 00:00:00 to 2016-09-13 17:00:00 since there was no previous data before this interval and 3) remove the duplicate timestamps that existed at the end of each month (explained in Chapter 3). The other action taken, which is presented in the framework, was the application of the moving average to fill the missing values presented in the data for either kWh consumption or weather related data.

With the data cleaned, the next step was to perform some data transformation, to create new features that could be important to explain the dependent variable which, in this case, is the kWh consumption. Since there is data regarding date values, six new columns were created: 1) “day_week”, 2) “working_day”, 3) “year”, 4) “month”, 5) “day” and 6) “hour”. The first column provides information regarding the day of the week, where encoding was done (value 0 for Monday, value 1 for Tuesday and so on, being the value 6 the last value, associated with Sunday). The second column indicates if it is a working day or not, by associating the value 1 and 0, respectively. The four last columns (“year”, “month”, “day” and “hour”) are the disaggregation of the Timestamp in specific time variables that will be used for the data analysis segment.

4.2. Data Analysis

During this section, the processed data described in section 4.1 is used to show the information that is presented to the user and identified in section 3.3.1.

The energy consumption data shows a clear seasonality, with no trend, presenting higher kWh values during the winter period (December, January and February) where it keeps decreasing until August. After August, the energy consumption starts increasing again until reaching its peak around December and January. Figure 6 shows the kWh consumption evolution as well as the kWh consumption after the application of the moving average method, highlighted by Winter, Summer and Autumn/Fall periods.

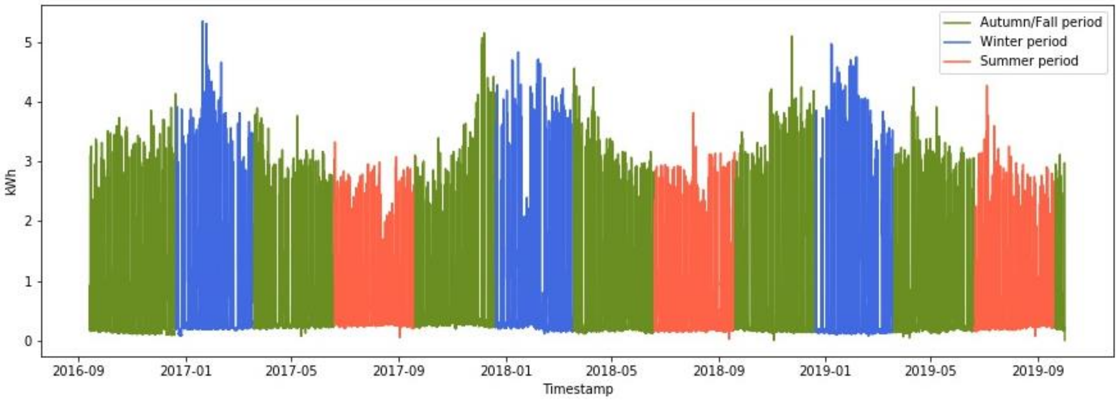

Figure 6 - Evolution of kWh consumption with season highlighting (after data treatment)

The cold temperatures that Portugal has during the winter period months, in conjunction with the building materials, can indicate that there was a need to use heating systems to keep all children and kindergarten assistants under a comfortable temperature, thus showing the high kWh consumption. Following the same logic in the summer period (June, July and August) where in some cases the temperature can reach 35º, it would be expected to also analyze an increase in energy consumption. The decrease in children at kindergarten can explain the observed lower kWh consumption during those months. Even though kindergarten vacation does not apply to the children of this kindergarten, only around 50% of children and kindergarten assistants are in kindergarten during this period, justifying the lower energy consumption. This type of information depends on a case by case, thus not being able to be delivered in the framework as a one size fits all.

Table 6, Table 7 and Table 8, provides information regarding average kWh consumption as well as max and min values recorded, top 5 values and bottom 5 values, with the respective timestamps.

Table 6 - Information on mean, min and max values for kWh consumption

Information	Value
Mean	0.64
Min	0.00
Max	5.36

Table 7 - Top 5 kWh consumption value

Timestamp	Value
2017-01-20 12:00:00	5.36
2017-01-24 12:00:00	5.32
2017-12-07 12:00:00	5.16
2017-01-20 13:00:00	5.13
2018-11-23 12:00:00	5.11

Table 8 - Bottom 5 kWh consumption value

Timestamp	Value
2018-11-03 06:00:00	0.00
2019-10-01 00:00:00	0.00
2018-11-03 03:00:00	0.01
2018-11-03 07:00:00	0.01
2018-11-03 08:00:00	0.01

Regarding the other variables, specifically “Temperatura (Cº)”, “Humidade (%)” and “Precipitação (mm)” by plotting the respective values through time, and as expected, it also shows seasonality with no clear trend. Also, when the variable “Temperatura (Cº)” increases the variables “Humidade (%)” and “Precipitação (mm)” decreases and vice-versa. These analyses can be confirmed in Appendix **A**, **B** and **C**.

By aggregating the value of energy consumption by the hour, day of the month, month and year, it also provides important information to identify the behavior in energy consumption, Using the data from the case study, it was shown that on the twelfth hour the cumulated energy consumption had the highest value, followed by the eleventh hour and the thirteenth hour, respectively. This could indicate that several appliances are being used at this time, for example, for lunch. Doing the same analysis for the other variables it shows that different days present higher values than other days, being the lowest the 31st day. The lower value on this day is explained by the difference in days of all twelve months, where only seven months have 31 days. For the other values, since there isn't enough information, it can be identified as a specific reason for the pattern. Regarding energy consumption by month, and as it was seen before, the month of January presents the highest values where the lowest value is associated with the month of June. For the month of January, it can be explained by the lowest temperatures observed during that month, as for the lowest value in June, the assumption is that it might be related with the increase in temperature in conjunction with the holidays that occur on the 10th and 13th of June, being the latter only associated with the district of Lisbon. Most families use these days to book vacations in between as a way to have more days for travelling, spend time with their families and relax. Regarding year, 2017 presents the highest value while 2016 has the lowest value. There isn't enough information to understand the higher value in 2017 as it may be explained by the increase in appliances, more people in kindergarten or noise in data. As for the lowest value in 2016, it is explained by the lack of records, since it begins only in September. Plots regarding kWh consumption by year, month, day of month and hour can be seen in appendix **D**, **E**, **F** and **G.**

Another information that is provided in the Data Analysis feature is the visualization of the correlation between the variables. The relationship between time-series variables can be analyzed through cross-correlation. Visually, the variables "kWh", "Temperatura (Cº)", "Humidade (%)" and "Precipitação (mm)" show a relationship in terms of evolution trough time, that would be indicative of the existence of some sort of correlation between them. The utilization of scatter plots between the dependent variable "kWh" and the variables "Temperatura (Cº)", "Humidade (%)" and "Precipitação (mm)" was generated, for this dataset, since it gives a previous visual identification of relationships between variables, where it was shown that there is a high dispersion of values in all three dependency relationship ("kWh" and "Temperatura (Cº)", "kWh" and "Humidade (%)", "kWh" and "Precipitação (mm)"). The generated scatter plots can be seen in appendix **H**, **I** and **J**.

The cross-correlation between these previously mentioned variables was analyzed, having a threshold of 0.6. This threshold, between 0 and 1, was defined with that value since a relationship of 0.6 would be considered good to conclude that the variable x was influenced by variable y. Figures 7, 8 and 9 show the cross-correlation between “kWh” and “Temperatura (Cº)”, “kWh” and “Humidade (%)” and “kWh” and “Precipitação (mm)”.

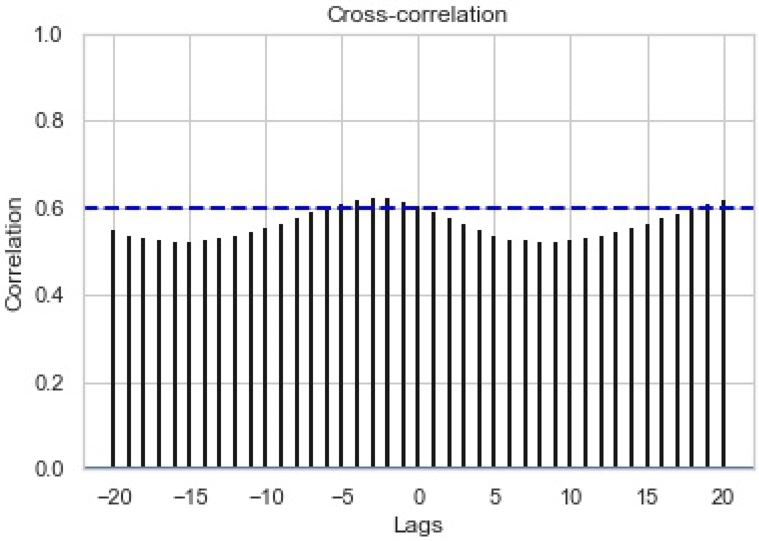

Figure 7 - Cross-correlation between “kWh” and “Temperatura (Cº)”

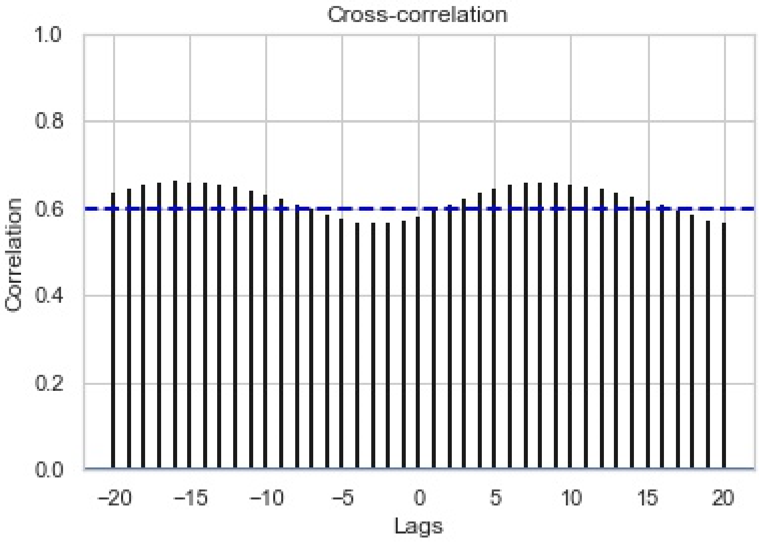

Figure 8 - Cross-correlation between “kWh” and “Humidade (%)”

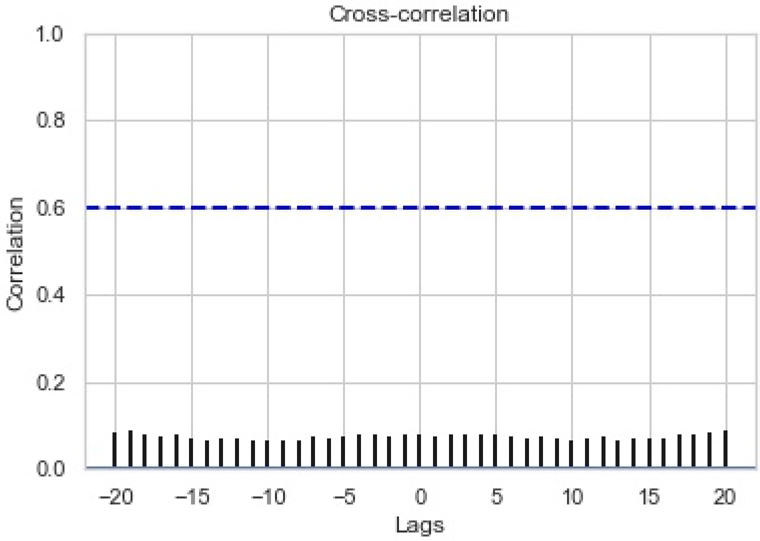

Figure 9 - Cross correlation between "kWh" and "Precipitação (mm)"

For "kWh" with "Temperatura (Cº)" and "kWh" with "Humidade (%)" some points in time get above the threshold line, hovering between 0.5 and 0.6. As for "kWh" and "Precipitation (mm)" the values only oscillate between 0.0 and 0.1. This led to the conclusion that these variables might not be good enough to help predict kWh consumption. However, for this case study, "Temperatura (Cº)" and "Humidade (%)" are used when testing multivariate predictive models in Chapter 5.

As for the other variables, the Pearson Correlation Coefficient values are presented in Figure 10.

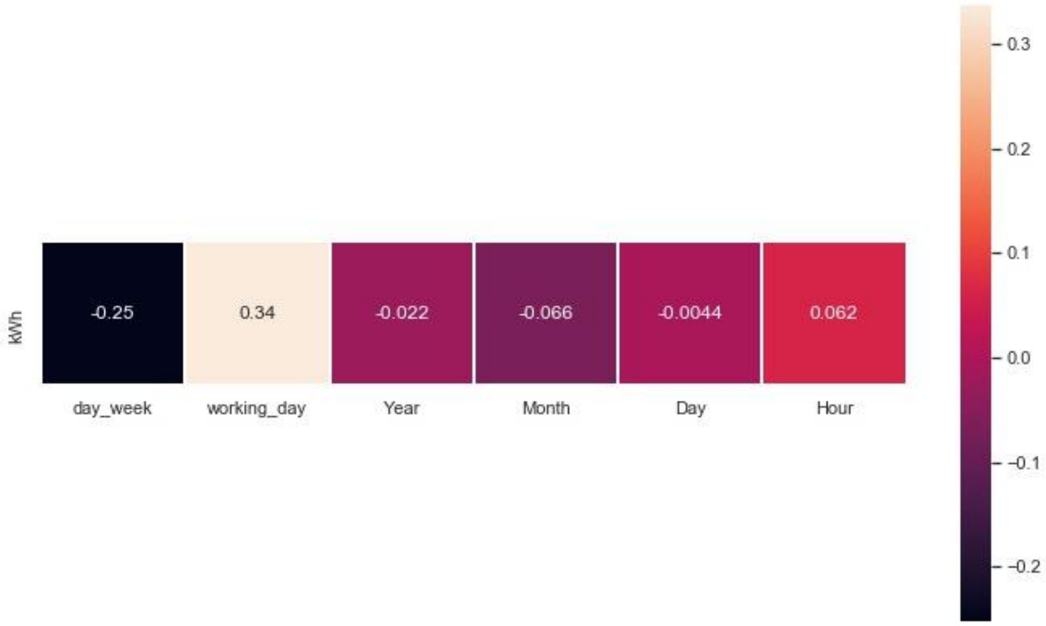

Figure 10 - Heatmap: Pearson correlation between variables

Analyzing the heatmap, the kWh column does not show any strong correlation between the other variables. The highest correlation value is 0.34 with the variable “working_day”, which is still considered a low correlation.

The aggregation of data regarding energy consumption by different time periods as well as the correlation between the variables are important for the identification of patterns and behaviors, that might help user take actions on different situations.

Chapter 5 – Predictive Models

Following the previous chapter, the focus of this chapter is the application of the predictive models for both situations, univariate and multivariate data, with a brief description of each predictive model, concluding on the best model to be used in the framework based on the data from the case study. For that, three models were selected and tested for each type of data, univariate and multivariate, taking into consideration the works analyzed in Chapter 2.

Other models could have been chosen for energy consumption forecasting based on the available works focused on this subject, however, one of the objectives is to have a forecasting stage that would use known and validated models to provide local approximate forecasted values in a simple and efficient way.

For that, the following steps were taken: 1) identification of some of the most commonly used predictive models for time-series problems based on literature review, 2) execution of the predictive models on the dataset, 3) results observed, 4) hyperparameter tuning, 5) re-evaluation, 6) model comparison and 7) model selection.

The predictive models applied to the dataset were developed on the open-source Python (version 3.6.8) with the usage of the following libraries:

- Pandas;
- Numpy;
- Matplotlib;
- Seaborn;
- Statsmodels;
- Keras;
- Scikit-Learn.

These libraries will enable functions for data treatment, predictive model development and data and results visualization.

5.1. Identification of the predictive models – Univariate

Through the analysis of published works related to energy consumption prediction, which has also been described in Chapter 2 of this document, some predictive models were selected to be applied to the kindergarten's dataset.

As previously stated, in time series, univariate data is described as only considering one variable for analysis. In other words, the variable itself contains the information needed to be forecasted. For this case study, the identification of the predictive models to be applied was based not only on the literature review done on Chapter 2 but also on the correct applicability for univariate data. From this analysis, three models were selected for further execution on the data.

5.1.1. SARIMA

The SARIMA model is a version of the ARMA model. The ARMA model is comprised by 2 sections: 1) AR (Autoregression) which is represented by "p" and MA (Moving Average) represented by "q". In case de data is not stationary, this model does not provide excellent results. The model is explained in (1).

$$Y_t = \mu + \sum_{j=1}^{p} \phi_j \left(Y_{t-j} - \mu\right) + \sum_{j=1}^{q} \psi_j \, \epsilon_{t-j} + \epsilon_t \tag{1}$$

To be able to deal with the trend in data, the integrated term "i" is applied, converting the not stationary data into stationary and naming the model ARIMA, being represented by the variable "d". To be able to take seasonality into account, the ARIMA model developed to a Seasonal ARIMA (SARIMA). The SARIMA model can be explained by (2), where "D" represents de seasonal differencing and "s" the seasonal period.

$$\Phi(B)\phi(B)\nabla_s^D \nabla^d Y_t = \Theta(B)\theta(B)Z_t \tag{2}$$

5.1.2. Holt-Winters Exponential Smoothing

The Holt-Winters Exponential Smoothing is a model that can perform well even if the data presents seasonal and trend variation. This model has two types of dealing with seasonality: 1) Additive and 2) Multiplicative. The additive Holt-Winters model takes into consideration that the seasonal variation is the same size throughout the periods and can be explained by (3).

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)} \tag{3}$$

The multiplicative Holt-Winters models consider that the seasonal variation changes in proportion with the time series in the analysis. This model can be explained by (4).

$$\hat{y}_{t+h|t} = (\ell_t + hb_t)\, s_{t+h-m(k+1)} \tag{4}$$

The “bt” expresses the trend component while “st” expresses the seasonal component.

5.1.3. Long Short-Term Memory

The Long Short-Term Memory Neural Network is a recurrent neural network, differentiating from the deep neural networks in terms of its usage.

The application of recurrent neural networks helps mitigate the loss of dependency by memorizing the important input throughout sequences. Although being able to hold information through the various sequences, it is still unable to hold that same information for extended sequences since the traditional recurrent neural networks utilize backpropagation in model training.

The LSTM was developed considering this loss and allowing the information to be held in longer sequences and is explained by (5) (from *a* to *f*).

$$i_t = \sigma(W_{xi}x_{[t]} + b_{xi} + W_{hi}h_{[t-1]} + b_{hi} \tag{5a}$$

$$f_t = \sigma(W_{xf}x_{[t]} + b_{xf} + W_{hf}h_{[t-1]} + b_{hf} \tag{5b}$$

$$g_t = tanh(W_{xg}x_{[t]} + b_{xg} + W_{hg}h_{[t-1]} + b_{hg} \tag{5c}$$

$$o_t = (W_{xo}x_{[t]} + b_{xo} + W_{ho}h_{[t-1]} + b_{ho} \tag{5d}$$

$$C_{[t]} = f_t \odot c_{[t-1]} + i_t \odot g_t \tag{5e}$$

$$h_{[t]} = o_t \odot tanh(c_{[t-1]}) \tag{5f}$$

The “i” indicates the input gate, “f” the forgotten information gate, “g” the update step, “o” the output gate, “c” the cell memory state and “h” the hidden state.

5.2. Identification of the predictive models – Multivariate

Multivariate data differentiates from univariate data by the number of variables presented on the dataset that can be feed to help models forecast our dependent variable. Due to this characteristic, some predictive models applied for univariate data cannot perform accurately when there is more than one variable that can help explain one specific variable. The following models were also analyzed through the literature review chapter, where it was scientifically proven that could be used for multivariate data or both (univariate and multivariate data).

5.2.1. SARIMAX

The SARIMAX model or Seasonal ARIMAX is similar to the seasonal ARIMA previously mentioned, with the inclusion of exogenous variables, defined by the X. The equation for the SARIMAX is explained by (6) [51].

$$\begin{aligned} Y_t = {} & B_0 + B_1 X_{1,t} + B_2 X_{2,t} + \cdots + B_k X_{k,t} \\ & + \frac{(1-\Theta_1 B-\Theta_2 B^2-\cdots-\Theta_q B^q)(1-\Theta_1 B^s-\Theta_2 B^s-\cdots--\Theta_Q B^{Qs})}{(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^q)(1-\Phi_1 B^s-\Phi_2 B^s-\cdots--\Phi_P B^{Qs})} \in_t \end{aligned} \tag{6}$$

5.2.2. Support Vector Regression

As stated in the literature review, the SVR is a variation of Support Vector Machines and is focused on minimizing the error forecasted on a training dataset while maintaining a functional form as flat as possible [52]. To allow the model to handle non-linear relationships, a Kernel function is provided. Some of the Kernel functions used for the model are Linear, Polynomial, Gaussian Radial Basis and Multilayer Perception. The equation for the SVR model is shown in equation (7).

$$f(x) = \sum_{i=1}^{N} (\beta_1^* - \beta_i) K(x_i - x) + b \tag{7}$$

5.2.3. Long Short-Term Memory

The LSTM model can also be used for forecasting with multivariate data. A brief explanation of this model has already been described in 5.1.3. in this document.

5.3. Evaluation Metrics

The application of these models in the dataset will predict values that will show errors, or in other words, it will show differences between the real and the predicted value. The value of the accuracy of these models depends on the forecast of these errors. Thus, for the evaluation of the models implemented in this work three measures of accuracy were used:1) Mean Absolute Error (MAE), 2) Mean Squared Error (MSE), 3) Root Mean Squared Error: (RMSE).

These measures can be explained by equation 8, 9 and 10, respectively.

$$MAE = mean(|e|) \tag{8}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \tag{9}$$

$$RMSE = \sqrt{mean(e_t^2)} \tag{10}$$

5.4. Application of the predictive models

This section presents the models' results for both univariate and multivariate data. Since the dataset used for this case study has data from three years, to correctly validate the models, it was applied cross-validation on a rolling basis, where the training sets will only consist on data that occurred prior to the data that exist in the test set [53]. Having three years of data, the data was sequentially divided into two sets, and each set was divided into training and test datasets. Table 9 shows how the data was divided and its characteristics.

Table 9 - Model Cross-Validation characteristics

Set	Training – Data Points	Test – Data Points
Set 1 (One year of data)	8761	17936
Set 2 (Two years of data)	17281	9416

5.4.1. Univariate Data

With the dataset only containing energy consumption by the hour, the predictive models applied were the previously identified as best suited for univariate data. All the selected models have different parameters that can be tuned to improve the quality of the model.

Specifically for the SARIMA models, there are too many different combinations that can be tested, such as combinations like (1,0,0) (0,0,0,24) or (1,1,0) (0,0,0,24), thus being very time consuming to manually apply each different combination in order to generate a model with the best evaluation metrics and that can accurately predict future values. There are other analysis that can be done to have a notion of the parameters that need to be applied in the SARIMA model, such as the Augmented Dickey-Fuller (ADF) test to identify if the data is stationary and the analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) to visually identify the values for the Autoregressive component (AR) and the Moving Average component (MA).

The application of the ADF test was applied to the full dataset, where the null hypothesis indicates that the data is not stationary. For the correct application of the SARIMA model, the objective is to reject the null hypothesis, meaning that the p-values must be lower than 0.05 where it was concluded that the data in the analysis was stationary (Test Statistic of -23.13 and a p-value of 0.00). This indicator also helps understand if there is the need to apply the differentiation in the data and consequently consider the differencing value into the “i” component in the SARIMA model.

In regards to the ACF and PACF, the ACF can be described as the linear relationship between two data points in time as a function of their time difference, the PACF is somewhat more complicated to understand, as it is described as the partial correlation of a specific lag is the partial correlation between the time series with itself at that lag will all the in-between information of the two points in time [17]. The ACF and PACF from the dataset are as follows:

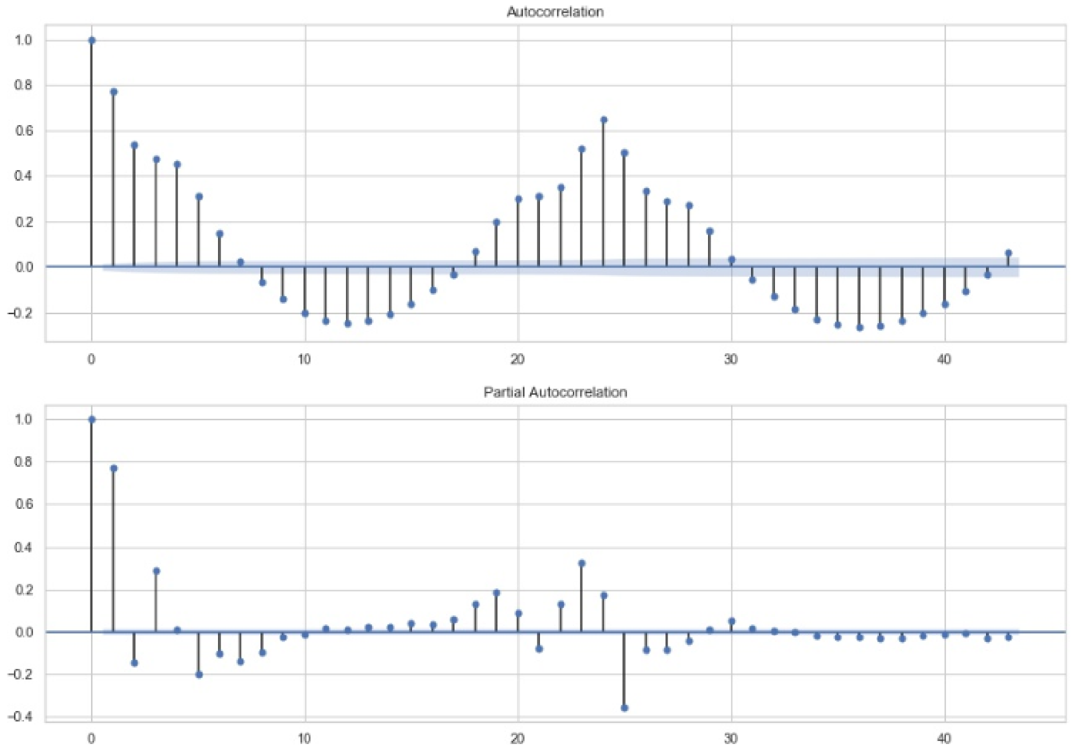

Figure 11 - ACF and PACF of the dataset only containing kWh consumption

From Figure 11 it can be seen in the ACF graph that there is a seasonality where peaks are observed in lag 0, 24, and considering the data in the analysis, it will also have peaks in lag 48, 72 and so on. As for the PACF, there are two peaks in lag 0 and 1. Both spikes at lag 1 are outside the significant level suggesting that there is a non-seasonal and seasonal component of 1 (MA (1)). Additionally, since seasonality is observed, the model will also have a seasonal differencing component of 1, thus for the first application of the SARIMA model the applied parameters were (0,0,1) (0,1,1) (24). Since the data was considered stationary, there was no need to apply the non-seasonal differencing.

For the Holt-Winters model, the additive method was considered for posterior analysis of the robustness of the model to forecast future values.

As for the LSTM model, the initial model was executed with the “ReLu” activation, 50 nodes, batch size of 20, one dropout layer, “adam” optimizer, “mse” loss and 30 epochs.

The following Table 10 provides the evaluation metrics of each model in each set when forecasting the next 24 hours:

Table 10 - Evaluation results from the univariate predictive models

Model / Evaluation Metric	Set	MSE	RMSE	MAE
SARIMA (0,0,1) (0,1,1,24)	1	0.39	0.62	0.36
Holt-Winters	1	0.17	0.42	0.38
LSTM	1	0.66	0.82	0.56
SARIMA (0,0,1) (0,1,1,24)	2	0.43	0.66	0.34
Holt-Winters	2	0.07	0.27	0.17
LSTM	2	0.74	0.86	0.47

It can be observed that the Holt-Winters model provided the best evaluation metrics for both sets, followed by the SARIMA model. By plotting the values for the next 24 hours using the second set of the Holt-Winters model, in Figure 12, it is visualized the accuracy of the forecasting by presenting very similar values between the real values and the predicted ones.

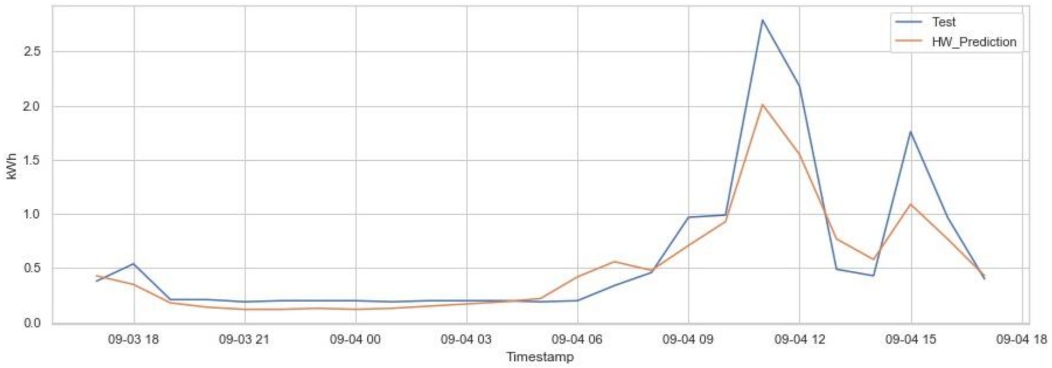

Figure 12 - Holt-Winters: Comparison between real and predicted values for the next 24 hours

5.4.2. Multivariate Data

Following the same logic as the univariate data, for the multivariate data, the dataset was divided first into two sets where each set was divided into train and test. The multivariate dataset considered only the variables “kWh”, “Temperatura (C^o)” and “Humidade (%)” because of the correlations observed between the variable we want to predict and the other variables.

For the SARIMAX model, the first execution used the same parameters for the univariate SARIMA model, with the addition of a training and test dataset only containing “Temperatura (Cº)” and “ Humidade (%)” to be later used in the exogenous component.

For the SVR model, the training and test dataset for both sets were divided into a dataset only containing the dependent variable and a dataset containing the independent variables, namely, one sample only considering “kWh” and the other sample “Temperatura (Cº)” and “Humidade (%)”. In terms of kernel the first execution used the “linear” method.

For the multivariate LSTM model, the first execution used the same parameters as the first execution of the univariate model.

Table 11 shows the results from the first execution of the multivariate predictive models for both sets.

Table 11 - Evaluation results from the multivariate predictive models

Model / Evaluation Metric	**Set**	**MSE**	**RMSE**	**MAE**
SARIMAX (0,0,1) (0,1,1,24)	1	0.39	0.62	0.38
SVR	1	0.74	0.86	0.46
LSTM	1	0.24	0.50	0.26
SARIMAX (0,0,1) (0,1,1,24)	2	0.43	0.66	0.34
SVR	2	0.74	0.86	0.45
LSTM	2	0.28	0.53	0.31

For the first multivariate execution, the LSTM model provided better accuracy results, for both sets, when compared with the SARIMAX and SVR.

By plotting the forecast for the next 24 hours (Figure 13), of the second set, and compare it to the real values, it can be shown that for the next 24 data points, the LSTM model did not provide reasonably accurate values but managed to understand the underlying trend of the data.

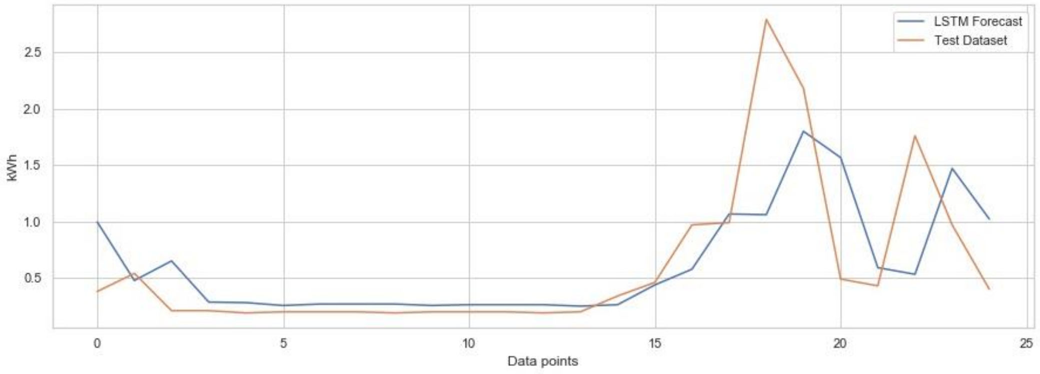

Figure 13 - LSTM: Comparison between real and predicted values for the next 24 hours

For both sets of the univariate and multivariate models, the evaluation results were very close to zero, providing the first indication, but not conclusive, that all models could be applied for predicting kWh consumption. In the next sections, hyperparameter tuning will be performed to analyze if the previous models can be improved and provided more accurate predicted values.

5.5. Model Optimization

Considering the results provided in the previous section 5.4. the next step was to optimize the models by adapting the hyperparameters used in the previous execution to understand if the models could be more accurate in terms of forecasting. Hyperparameter optimization is a very important step for machine learning since it can improve the results provided by the models instead of having the necessity to defining new learning paradigms [54].

Starting with the univariate data, from the three previously selected models, the SARIMA and LSTM models provide the highest applicability for hyperparameter tuning. As for the Holt-Winters, only the seasonal method can be changed. The high combination of non-seasonal and seasonal parameters in the SARIMA model makes it very time consuming to manually test each scenario in order to identify the best results. For the LSTM, the hyperparameter tuning is applied to the number of inputs, hidden layers, dropout layers, activation method, optimizer method and loss method. Despite these actions for improving the evaluation metrics of the models it is also important to refer that the data forecasted also needs to be visually analyzed in order to confirm if the models can fairly predict the values or classes we want to predict.

The *auto_arima* library, available in Python, was used for the analysis of different combinations that we want the SARIMA model to run on and give the Aikaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC) value of each combination. The AIC and BIC are criteria methods that can be used to compare two models as a way to help for model selection. The objective of the AIC is to select the model that presents the most negative likelihood, penalized by the number of parameters [55]. The AIC is explained by (9):

$$AIC = -2\log p(L) + 2p \tag{11}$$

As for the BIC it follows the same logic as the AIC only differing on the second term that depends on sample size [55]. The BIC is explained by (10):

$$AIC = -2\log p(L) + p\,log\,(n) \tag{12}$$

Having into account the SARIMA model and the parameters that influence the robustness and the applicability for prediction, the *auto_arima* process was run considering a set of values as the intervals of analysis. The intervals are presented in Table 12.

Table 12 - SARIMA parameters intervals for auto_arima

SARIMA Parameters	**Intervals**
p (AR)	[0,3]
d (Differencing)	0
q (MA)	[0,3]
P (Seasonal AR)	[0,3]
D (Seasonal Differencing)	0
Q (Seasonal MA)	[0,3]

Since the stationarity of the data was already confirmed by the ADF test, the value for the non-seasonal and seasonal differencing component was maintained as 0. The *auto_arima* process delivered a list of combinations with the respective AIC and BIC values, which a few can be seen in Table 13.

Table 13 - AIC and BIC values from different SARIMA combinations

SARIMA Model Values	Set	AIC	BIC
SARIMA (1,0,0) x (0,1,0,24)	1	9985.22	10006.44
SARIMA (2,0,1) x (2,1,0,24)	1	8779.38	8779.38
SARIMA (3,0,3) x (3,1,1,24)	1	6046.49	6131.39
SARIMA (3,0,3) x (2,1,1,24)	1	6166.08	6243.91
SARIMA (3,0,2) x (3,1,1,24)	1	6102.51	6180.34
SARIMA (1,0,0) x (1,1,0,24)	2	19498.08	19529.11
SARIMA (2,0,1) x (3,1,1,24)	2	13552.26	13622.07
SARIMA (1,0,0) x (3,1,1,24)	2	13603.26	13657.55
SARIMA (2,0,1) x (3,1,0,24)	2	17495.52	17542.05
SARIMA (2,0,1) x (2,1,1,24)	2	13742.30	13804.34

For SARIMA, both sets have different AIC and BIC values, due to the number of data points presented in each step. For the first set, the application of the parameter (3,0,3) x (3,1,1,24) provided the lowest AIC and BIC value. As for the second set, the application of the parameters (2,0,1) x (3,1,1,24) showed the lowest value of the AIC and BIC. These results were obtained, considering the intervals identified in Table 11.

For the Holt-Winters model, the initial additive method was changed to the multiplicative method.

Lastly, for the LSTM model, the number of nodes was changed to 200, the batch size to 30 and the inclusion of a second layer and a second dropout. The dropout helps prevent overfitting in removing inputs to a layer from a previous layer.

Moving to the multivariate data, for the SARIMAX with exogenous variables, the parameter identification will follow the same logic as the SARIMA with univariate data. The usage of the *auto_arima* with the addition of an "exogenous" component helped in analyzing the best AIC values for a multivariate dataset. In terms of intervals, the same values presented in Table 11 were used. Table 14 presents some of the AIC and BIC values after the execution of the *auto_arima* considering exogenous variables.

Table 14 - AIC and BIC values from different SARIMAX combinations

SARIMA Model Values	Set	AIC	BIC
SARIMA (1,0,1) x (3,1,1,24)	1	6446.25	6517.00
SARIMA (1,0,0) x (3,1,1,24)	1	6454.08	6517.76
SARIMA (3,0,1) x (3,1,1,24)	1	6162.43	6247.34
SARIMA (3,0,1) x (2,1,1,24)	1	6307.19	6385.02
SARIMA (1,0,0) x (2,1,1,24)	1	6600.94	6657.54
SARIMA (1,0,0) x (1,1,1,24)	2	14187.95	14242.24
SARIMA (1,0,1) x (1,1,1,24)	2	14191.22	14253.26
SARIMA (1,0,2) x (1,1,1,24)	2	13536.32	13606.12
SARIMA (3,0,2) x (2,1,2,24)	2	12980.01	13080.84
SARIMA (2,0,2) x (3,1,1,24)	2	13044.73	13137.80

For the SVR model, prior to using the “linear” kernel component, other methods were used, namely “poly”, “rbf” and “sigmoid”. Table 15 presents the R^2 values obtained for the SVR when running all methods through the data.

Table 15 - Score values using different methods for SVR

Method	Set	R^2
linear	1	-0.14
rbf	1	0.02
poly	1	-946372778.73
sigmoid	1	-0.18
linear	2	-0.17
rbf	2	-0.15
poly	2	-0.13
sigmoid	2	-68234.42

All the R^2 values, besides the R^2 from the "rbf" method of set 1, presented a negative value, which is an indication that the predictive model may not be suited for our purpose. The new executions for both sets used the "rbf" method.

Lastly, for the LSTM model, the same parameters altered for the univariate data were used in the multivariate data.

5.6. Model reevaluation

After executing the models again, with the new parameters, the following evaluation metrics for the univariate models were generated, as presented in Table 16.

Table 16 - Univariate: Evaluation results from the predictive models after hyperparameter tuning

Model / Evaluation Metric	**Set**	**MSE**	**RMSE**	**MAE**
SARIMA (3,0,3) x (3,1,1,24)	1	0.36	0.60	0.41
Holt-Winters	1	0.06	0.25	0.21
LSTM	1	0.42	0.65	0.32
SARIMA (2,0,1) x (3,1,1,24)	2	0.36	0.60	0.32
Holt-Winters	2	0.17	0.40	0.22
LSTM	2	0.73	0.85	0.50

Analyzing the three new executions for both sets, the Holt-Winters model continued showing the best results, although set 2 presented worse results that its first execution using the additive method. Overall, the SARIMA and LSTM models presented better values when compared with its previous execution. Figure 14 shows the forecast for the next 24 hours using Holt-Winters models and compares them with real values.

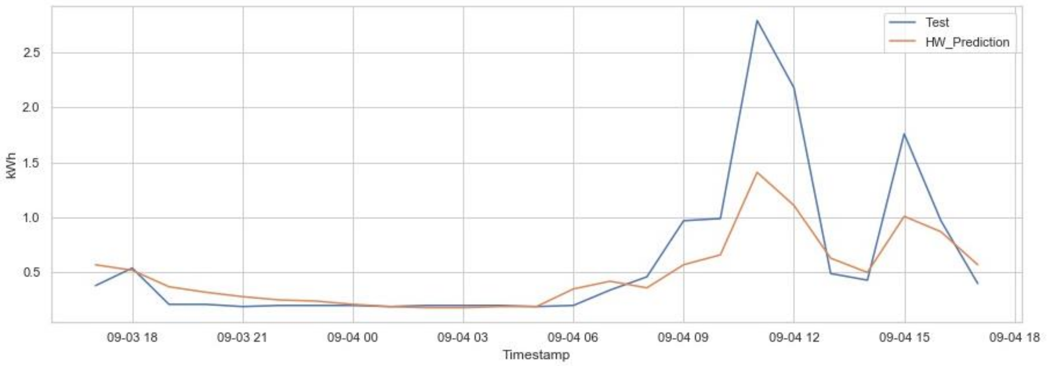

Figure 14 - Holt-Winters: Comparison between real and predicted values for the next 24 hours

The graph shows the Holt-Winters model accurately forecasted the daily seasonality of the data, although showing a higher error between the forecasted and real values during the peak hours.

As for the multivariate data, the new executions provided the following results as presented in Table 17.

Table 17 - Multivariate: Evaluation results from the predictive models after hyperparameter tuning

Model / Evaluation Metric	Set	MSE	RMSE	MAE
SARIMAX (3,0,1) x (3,1,1,24)	1	0.13	0.37	0.28
SVR	1	0.72	0.85	0.48
LSTM	1	0.41	0.65	0.32
SARIMAX (3,0,2) x (2,1,2,24)	2	0.07	0.26	0.16
SVR	2	0.71	0.84	0.47
LSTM	2	0.37	0.60	0.32

Both sets for the SARIMAX model provided better results when compared with the SVR and LSTM. For the SVR, it was projected that the model would not make very good predictions by analyzing the negative R^2 scores it provided. This LSTM model got worse results when compared to its first execution.

Figure 15 shows the forecast for the next 24 hours using SARIMAX model, using set 2, where it can be observed that the model generated accurate forecasted kWh values, when compared to test set.

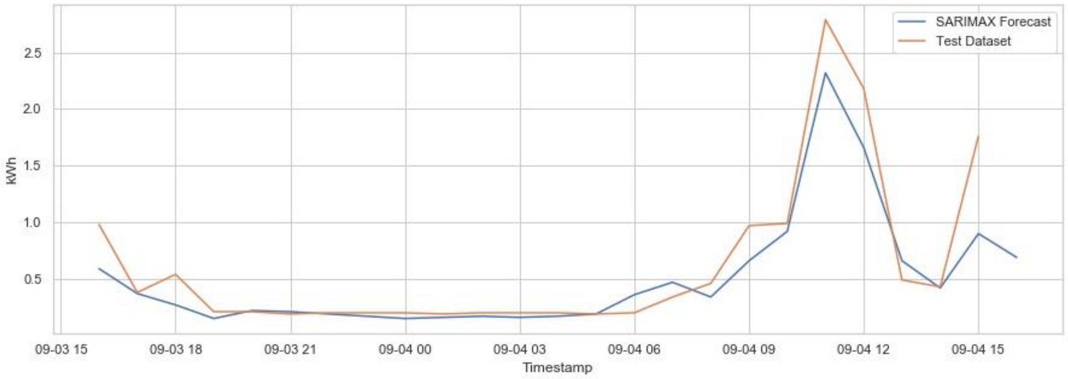

Figure 15 - SARIMAX: Comparison between real and predicted values for the next 24 hours

By analyzing the evaluation metrics, all models presented values close to 0 and always lower than 1, implying that they could be applied to the framework to forecast kWh consumption values. However, when plotting the comparison between the real and the forecasted values of some models, it can be seen a high disparity between them. Figure 16 shows the 24 hours forecast provided by the latter execution of the LSTM model.

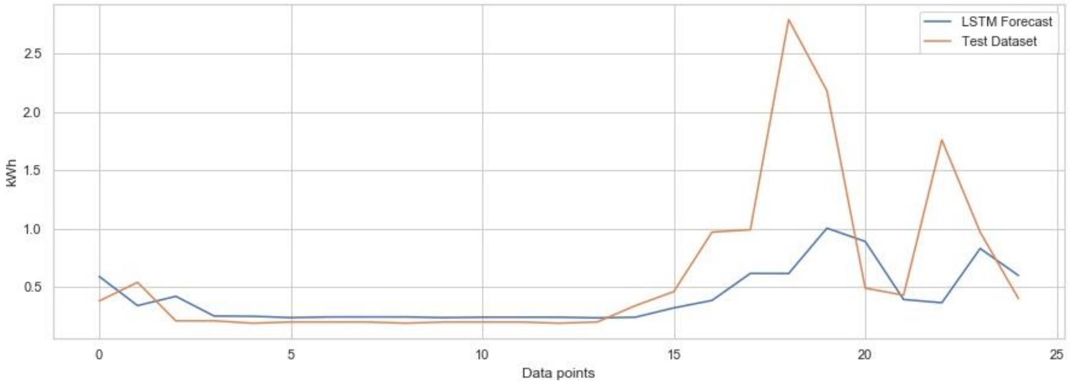

Figure 16 - LSTM: Comparison between real and predicted values for the next 24 hours

To sum up, the hyperparameter tuning of the predictive models can provide better or worst results than the first execution as it can be seen for the Holt-Winter models. Nonetheless, it is important to plot the data and see the real comparison between real and forecasted values. In Figure 16, it can be seen that the model identified the seasonality of the data, however the error between the real value and the predicted one is very high during the hours when kWh consumption increases. By analyzing the evaluation metrics, we would assume that the model was good enough to forecast kWh consumption, however with visualization techniques, we can conclude on the viability of the model for forecasting.

5.7. Predictive model selection

Additionally, to the evaluation metrics generated in section 5.6 for both univariate and multivariate data, two other factors were taken into account to decide on the predictive models used for the framework:

- Execution time;
- Optimization.

5.7.1. Execution time

As stated in 5.5, from the selected model for testing, some had more parameters to be tuned that can help the model to better forecast values.

The SARIMA models (both univariate and multivariate) took between two to five minutes to execute, depending on the parameters selected. However, these models have a high number of possible combinations that will take time until the best model is identified. Even with the application of methods, such as the *auto_arima* code used in this work with different intervals, the execution time finished around 23 thousand seconds (6 hours) which is far too long. The *auto_arima* code with high *m* values (in this case, 24) makes the seasonal ARIMA processes to take much time processing.

The Holt-Winters provided the lowest time executions (around two seconds).

The LSTM model will depend on the number of nodes, hidden layers, batch size and epochs, being that for the case study each batch was concluding between six to seven seconds.

Lastly, the SVR model also provided very low execution times of around three to four seconds.

5.7.2. Optimization

The meaning of optimization in this section relates to the scalability for hyperparameter tuning.

Both univariate and multivariate SARIMA models have several parameter combinations that will provide better or worse results depending on the values selected for those parameters. Visually we can identify the parameter values by analyzing the ACF and PACF of the data, but by analyzing the AIC and BIC value, we can compare between different parameter combinations and select the best one.

The Holt-Winters and SVR models have fewer optimization options, being the method used in each model the only parameter that can really have an impact on the accuracy of the models.

The LSTM models can be modelled with different nodes, hidden layers, dropout layers, batch sizes, epochs, function activation, optimizers and function losses, so at par with the SARIMA models, the LSTM also has high scalability for parameter tuning.

5.7.3. Model selection

The selection of the predictive models for both univariate and multivariate data was based on the evaluation metrics observed in section 5.4 and 5.6 and on the two factors explained in section 5.7. with the conclusion that SARIMA model was most suited for the framework, for both univariate and multivariate data.

Although it was stated that the SARIMA models take much time when analyzing the best parameters to have the lower AIC this could be mitigated by having more processing power. This action only needs to be done for new implementations or for updating the process after new data has been collected that shows new characteristics (for example, improving building characteristics to be more energy-efficient or buying more energy-efficient appliances). Also, by having many combinations for the different parameters, it is assumed that it can also adapt to different scenarios and kWh consumption behaviors from different buildings or households and still provide viable approximate local kWh consumption forecasts.

Chapter 6 – Conclusions

The purpose of the work was the conceptualization of a framework that could provide information regarding kWh consumption and approximate local forecasts in buildings or households. This framework could help end-users take actions by analyzing their energy consumption footprint and identify possible anomalies that might need to be addressed. While this proposed framework is mainly designed to be applied to each case individually, by scaling it to a more connected network, it can be used by municipalities to understand what energy policies should be promoted to improve energy efficiency and reduce energy waste and CO2 emissions.

Initially, a methodology was identified for the definition of the framework. The methodology selected was the widely known and industry-accepted CRISP-DM which provides six stages of actions related with problem definition, modelling and evaluating results which were used in work to help conceptualize and experiment some components of the proposed framework. The data collected from a private kindergarten was used not only to test some steps of the framework but also to understand the data requirements necessary for the conceptualization of the framework to be applied in different buildings or households.

Although the framework was conceptualized as an end-to-end process, the instalment and analysis of the IoT systems that would extract the data were not described in work.

The kWh recorded, from September 2016 to September 2019, was related to the general utilization of energy in the kindergarten, without any disaggregation by room, appliances or other energy consumption variables. Also, this data takes into consideration the energy losses that occur from the architectural and materials used for the building that might retain different thermal conditions (for example, rooms maintain energy and stay warmer in the winter or cooler in the summer). Additionally, to the kWh consumption, data regarding temperature, humidity and precipitation was also extracted from the wunderground.com website, since it provided historical hourly data that could be correlated with the kWh data. The utilization of these data helped conclude on the applicability of the framework for the defined investigation question as well as on the predictive model selection.

When analyzing the data quality from the kindergarten data, it was identified that, for the framework, the application of the moving average method to fill the missing values was the most suited one, while outliers were considered as having an important aspect for anomaly detection during the data analysis. As for the correlation between kWh consumption and other variables, temperature and humidity showed a good correlation, while precipitation presented a very low value. The reason for precipitation not having a strong correlation is identified by the lack of influence it has on the thermal ambient of the divisions. Also, other created variables, such as day of the month, month, year, working day and hour, did not show a strong Pearson correlation value. Additionally, since the kindergarten is open even during holidays, there is not a specific drop on energy consumption during a specific period, maintaining a stable wave line shape throughout the three years of data.

Having the data analysis performed, the next step was to test the predictive models. The data was divided into two scenarios, univariate data which only comprised of kWh consumption and multivariate data, which comprised of kWh consumption, "Temperatura (Cº)" and "Humidade (%)".

Three predictive models were used for both univariate and multivariate data. For the univariate data, a seasonal ARIMA (SARIMA), Holt-Winters and a recurrent neural network LSTM model were used. As for the multivariate data, the seasonal ARIMA and LSTM were also used, with the addition of the inclusion of exogenous variables (in this case "Temperatura (Cº)" and "Humidade (%)"). The Holt-Winter did not perform correctly with this type of data, so a version of the Support Vector Machines was used. This model is defined as Support Vector Regression (SVR).

Regarding forecasting for the univariate data, with the first execution and after the model optimization, the Holt-Winters provided the best evaluation metrics, with MSE, RMSE and MAE of 0.07, 0.27 and 0.17, respectively. Despite these results, in terms of adaptability for the different cases, the SARIMA model was selected as the model to used in the framework, which also provided very good evaluation results. Visually the SARIMA model, shows good results when comparing the real values with the forecasted ones.

As for the multivariate data, the SARIMA model with the parameters (3,0,2) x (2,1,2,24) provided the best MSE, RMSE and MAE values, with 0.07, 0.26 and 0.16.

This model was also selected to be used for multivariate data and visually also showed good results when comparing the real values with the forecasted ones.

Despite having concluded on the applicability of the framework for data treatment, data analysis and forecasting, it was not materialized into a running prototype that would incorporate the Python code with a functional software that would allow for the demonstration of the Data Analysis and Predictive Models features as they were conceptualized.

6.1. Framework applicability

The Data Analysis and Predictive Models features described and tested in this work helped understand that the development of an energy consumption forecast framework that would provide approximate local forecasts is viable. By developing a framework that follows a standard on collecting data, prepare it and create insights and forecasts on kWh consumption enables its applicability for several scenarios. For this work, the framework was tested with data collected from a kindergarten, but it could also be applied to domestic households, faculties or companies that see the necessity and importance on having real-time information of future kWh consumption to, for example, decrease costs.

The application of the predictive models for the univariate data was also applied in the ISCTE project: Social_IoT - University Community Engagement in Technologies for Sustainability: a Social Architecture. This work is part of a project published a paper in the scientific paper in a Quartil one journal of Applied Sciences ((ISSN 2076-3417 and impact factor of 2.74): Mataloto, B., Mendes, H. & Ferreira, J. (2020). Things2People Interaction Towards Energy Savings in Shared Spaces using BIM.

6.2. Future Work

Generally, the proposed framework could have some improvements to be made. In this work, the framework only described with detail steps and actions regarding data treatment, data analysis and forecasting, where data as already been collected from a source. As a first improvement, this framework could also describe with more detail the data extraction component before entering the database, namely IoT systems and the web scrapping techniques to gather kWh, temperature and weather data.

Next, and as stated in section 3.3, it would be more useful for this framework to be implemented by an energy provider since it would also give them general information regarding clients energy consumption and forecast values that would help identify the needed energy to provide to the grid. Also, having the contracted price with the clients, they could provide correct prices for the forecasted values.

For data analysis, other types of analysis could be implemented based on user feedback for what they would like to analyze regarding their kWh consumption, such as thresholds on energy consumption that would send alerts to the users. Also, the inclusion of more detail for data treatment techniques for daily and monthly data would provide more empirical evidence on the applicability of the predictive models for different data aggregations and user necessities.

Regarding the predictive models, other models could have been tested with the possibility to identify a better predictive model with different data aggregations (for example, daily data). Also, the process could consider more than one predictive model per data type (univariate and multivariate) that, in conjunction with a weighting system to identify the best predictive model for that specific case, could enable the system do adapt to the different specificities of each building or household and data aggregations.

Lastly, the real development of processes and software that would incorporate all the framework features would provide better conclusions on its applicability, by demonstrating execution times and visualizations regarding the information provided by the data analysis techniques as well as forecasted values.

References

[1] BPSTATS, “BP Statistical Review of World Energy Statistical Review of World, 68th edition,” *Ed. BP Stat. Rev. World Energy*, pp. 1–69, 2019. [Online]. Available: https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats review-2019-full-report.pdf. [Acessed: July, 11th 2020]

[2] T. Covert, M. Greenstone, and C. R. Knittel, “Will We Ever Stop Using Fossil Fuels?,” *J. Econ. Perspect.*, vol. 30, no. 1, pp. 117–138, Feb. 2016, doi: 10.1257/jep.30.1.117.

[3] R. A. Barreto, “Fossil fuels, alternative energy and economic growth,” *Econ. Model.*, vol. 75, no. July, pp. 196–220, 2018, doi: 10.1016/j.econmod.2018.06.019.

[4] IRENA International Renewable Energy Agency, *Renewable Power Generation Costs in 2017*. 2018.

[5] C. Jin, X. Sheng, and P. Ghosh, “Optimized electric vehicle charging with intermittent renewable energy sources,” *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 6, pp. 1063–1072, 2014, doi: 10.1109/JSTSP.2014.2336624.

[6] T. Kousksou, P. Bruel, A. Jamil, T. El Rha, and Y. Zeraouli, “Solar Energy Materials & Solar Cells Energy storage : Applications and challenges,” vol. 120, pp. 59–80, 2014, doi: 10.1016/j.solmat.2013.08.015.

[7] N. Brennan and T. M. Van Rensburg, “Wind farm externalities and public preferences for community consultation in Ireland: A discrete choice experiments approach,” *Energy Policy*, vol. 94, pp. 355–365, 2016, doi: 10.1016/j.enpol.2016.04.031.

[8] M. Santamouris, “Innovating to zero the building sector in Europe: Minimising the energy consumption, eradication of the energy poverty and mitigating the local climate change,” *Sol. Energy*, vol. 128, pp. 61–94, 2016, doi: 10.1016/j.solener.2016.01.021.

[9] C. Deb and S. E. Lee, “Determining key variables influencing energy consumption in office buildings through cluster analysis of pre- and post-retrofit building data,” *Energy Build.*, vol. 159, pp. 228–245, 2018, doi: 10.1016/j.enbuild.2017.11.007.

[10] A. Jozi, T. Pinto, I. Praça, F. Silva, B. Teixeira, and Z. Vale, “Wang and

Mendel’s fuzzy rule learning method for energy consumption forecasting considering the influence of environmental temperature,” *2016 Glob. Inf. Infrastruct. Netw. Symp. GIIS 2016*, 2017, doi: 10.1109/GIIS.2016.7814944.

[11] R. Zhang and H. Yang, “Dynamic building energy consumption forecast using weather forecast interpolations,” *2015 IEEE Int. Conf. Smart Grid Commun. SmartGridComm 2015*, pp. 671–676, 2016, doi: 10.1109/SmartGridComm.2015.7436378.

[12] J. G. De Gooijer and R. J. Hyndman, “25 Years of Time Series Forecasting,” *Int. J. Forecast.*, vol. 22, no. 3, pp. 443–473, 2006, doi: 10.1016/j.ijforecast.2006.01.001.

[13] M. Labriet, S.R. Joshi, F. Babonneau, N. R. Edwards, P. B. Holden, A. Kanudia, R. Loulou and M. Vielle, “Worldwide impacts of climate change on energy fora heating and cooling,” *Mitig. Adapt. Strateg. Glob. Chang.*, vol. 20, no. 7, pp. 1111–1136, 2015, doi: 10.1007/s11027-013-9522-7.

[14] ERSE, “Resumo Informativo - Comparação preços eurostat,” 2019. [Online]. Available: https://www.apren.pt/contents/publicationsothers/comparacao-precos-eurostat-dezembro-2017-erse.pdf. [Acessed: July, 11th 2020]

[15] Eurostat, “Household expenditure in 2018,” 2019. Available: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/DDN-20191127-1. [Acessed: September, 15th 2020]

[16] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, “Crisp-Dm 1.0,” *CRISP-DM. Consortium.*, p. 76, 2000. [Online]. Available: https://the-modeling-agency.com/crisp-dm.pdf [Acessed: March, 20th 2020]

[17] A. Nielsen, *Practical Time Series Analysis Prediction with Statistics & Machine Learning*, First Edit. O´Reilly, 2019.

[18] L. Suganthi and A. A. Samuel, “Energy models for demand forecasting - A review,” *Renew. Sustain. Energy Rev.*, vol. 16, no. 2, pp. 1223–1240, 2012, doi: 10.1016/j.rser.2011.08.014.

[19] A. Rahman, V. Srikumar, and A. D. Smith, “Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks,” *Appl. Energy*, vol. 212, no. December 2017, pp. 372–385, 2018, doi: 10.1016/j.apenergy.2017.12.051.

[20] C. Yuan, S. Liu, and Z. Fang, “Comparison of China’s primary energy consumption forecasting by using ARIMA (the autoregressive integrated moving average) model and GM(1,1) model,” *Energy*, vol. 100, pp. 384–390, 2016, doi: 10.1016/j.energy.2016.02.001.

[21] P. Sen, M. Roy, and P. Pal, “Application of ARIMA for forecasting energy consumption and GHG emission: A case study of an Indian pig iron manufacturing organization,” *Energy*, vol. 116, pp. 1031–1038, 2016, doi: 10.1016/j.energy.2016.10.068.

[22] C. Nichiforov, I. Stamatescu, I. Fagarasan, and G. Stamatescu, “Energy consumption forecasting using ARIMA and neural network models,” *Proc. - 2017 5th Int. Symp. Electr. Electron. Eng. ISEEE 2017*, vol. 2017-Decem, pp. 1–4, 2017, doi: 10.1109/ISEEE.2017.8170657.

[23] A. Rahman and A. S. Ahmar, “Forecasting of primary energy consumption data in the United States: A comparison between ARIMA and Holter-Winters models,” *AIP Conf. Proc.*, vol. 1885, no. September, 2017, doi: 10.1063/1.5002357.

[24] A. Hussain, M. Rahman, and J. A. Memon, “Forecasting electricity consumption in Pakistan: The way forward,” *Energy Policy*, vol. 90, pp. 73–80, 2016, doi: 10.1016/j.enpol.2015.11.028.

[25] S. I. Vagropoulos, G. I. Chouliaras, E. G. Kardakos, C. K. Simoglou, and A. G. Bakirtzis, “Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting,” *2016 IEEE Int. Energy Conf. ENERGYCON 2016*, 2016, doi: 10.1109/ENERGYCON.2016.7514029.

[26] K. Jeong, C. Koo, and T. Hong, “An estimation model for determining the annual energy cost budget in educational facilities using SARIMA (seasonal autoregressive integrated moving average) and ANN (artificial neural network),” *Energy*, vol. 71, pp. 71–79, 2014, doi: 10.1016/j.energy.2014.04.027.

[27] A. Sarhadi, R. Kelly, and R. Modarres, “Snow water equivalent time-series forecasting in Ontario, Canada, in link to large atmospheric circulations,” *Hydrol. Process.*, vol. 28, no. 16, pp. 4640–4653, 2014, doi: 10.1002/hyp.10184.

[28] F. Taşpinar, N. Çelebi, and N. Tutkun, "Forecasting of daily natural gas consumption on regional basis in Turkey using various computational methods," *Energy Build.*, vol. 56, pp. 23–31, 2013, doi: 10.1016/j.enbuild.2012.10.023.

[29] O. Kramer and F. Gieseke, "Short-term wind energy forecasting using support vector regression," *Adv. Intell. Soft Comput.*, vol. 87, pp. 271–280, 2011, doi: 10.1007/978-3-642-19644-7_29.

[30] G. Oğcu, O. F. Demirel, and S. Zaim, "Forecasting Electricity Consumption with Neural Networks and Support Vector Regression," *Procedia - Soc. Behav. Sci.*, vol. 58, pp. 1576–1585, 2012, doi: 10.1016/j.sbspro.2012.09.1144.

[31] U. K. Das, K. S. Tey, M. Seyedmahmoudian, M. Y. I. Idris, S. Mekhilef, B. Horan and A. Stojcevski, "SVR-based model to forecast PV power generation under differentweather conditions," *Energies*, vol. 10, no. 7, pp. 1–17, 2017, doi: 10.3390/en10070876.

[32] K. Kandananond, "Forecasting electricity demand in Thailand with an artificial neural network approach," *Energies*, vol. 4, no. 8, pp. 1246–1257, 2011, doi: 10.3390/en4081246.

[33] O. A. Olanrewaju, A. A. Jimoh, and P. A. Kholopane, "Comparing performance of MLP and RBF neural network models for predicting South Africa's energy consumption," *J. Energy South. Africa*, vol. 23, no. 3, pp. 40–46, 2012.

[34] H. T. Pao, "Comparing linear and nonlinear forecasts for Taiwan's electricity consumption," *Energy*, vol. 31, no. 12, pp. 2129–2141, 2006, doi: 10.1016/j.energy.2005.08.010.

[35] H. X. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 16, no. 6, pp. 3586–3592, 2012, doi: 10.1016/j.rser.2012.02.049.

[36] H. Musbah and M. El-Hawary, "SARIMA Model Forecasting of Short-Term Electrical Load Data Augmented by Fast Fourier Transform Seasonality Detection," *2019 IEEE Can. Conf. Electr. Comput. Eng. CCECE 2019*, pp. 1–4, 2019, doi: 10.1109/CCECE.2019.8861542.

[37] A. Blázquez-García, A. Conde, A. Milo, R. Sánchez, and I. Barrio, "Short-term office building elevator energy consumption forecast using SARIMA," *J. Build. Perform. Simul.*, vol. 13, no. 1, pp. 69–78, 2020, doi:

10.1080/19401493.2019.1698657.

[38] F. Yasmeen and M. Sharif, “Forecasting Electricity Consumption for Pakistan,” *Int. J. Emerg. Technol. Adv. Eng.*, vol. 4, no. 4, pp. 496–503, 2014.

[39] J. A. Zancanaro, O. L. Dos Santos, L. F. Ugarte, M. Giesbrecht, and M. C. De Almeida, “Energy Consumption Forecasting Using SARIMA and NARNET: An Actual Case Study at University Campus,” *2019 IEEE PES Conf. Innov. Smart Grid Technol. ISGT Lat. Am. 2019*, pp. 1–6, 2019, doi: 10.1109/ISGT-LA.2019.8895323.

[40] Y. Chen, P. Xu, Y. Chu, W. Li, Y. Wu, L. Ni. Y. Bao and K. Wang., “Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings,” Appl. Energy, vol. 195, pp. 659–670, 2017, doi: 10.1016/j.apenergy.2017.03.034.

[41] D. Liu, Q. Chen, and K. Mori, “Time series forecasting method of building energy consumption using support vector regression,” *2015 IEEE Int. Conf. Inf. Autom. ICIA 2015 - conjunction with 2015 IEEE Int. Conf. Autom. Logist.*, no. August, pp. 1628–1632, 2015, doi: 10.1109/ICInfA.2015.7279546.

[42] M. W. Ahmad, M. Mourshed, and Y. Rezgui, “Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption,” *Energy Build.*, vol. 147, pp. 77–89, 2017, doi: 10.1016/j.enbuild.2017.04.038.

[43] Y. T. Chae, R. Horesh, Y. Hwang, and Y. M. Lee, “Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings,” *Energy Build.*, vol. 111, pp. 184–194, 2016, doi: 10.1016/j.enbuild.2015.11.045.

[44] J. Yuan, C. Farnham, C. Azuma, and K. Emura, “Predictive artificial neural network models to forecast the seasonal hourly electricity consumption for a University Campus,” *Sustain. Cities Soc.*, vol. 42, no. June, pp. 82–92, 2018, doi: 10.1016/j.scs.2018.06.019.

[45] L. G. B. Ruiz, M. C. Pegalajar, M. Molina-Solana, and Y. K. Guo, “A case study on understanding energy consumption through prediction and visualization (VIMOEN),” *J. Build. Eng.*, vol. 30, no. February, p. 101315, 2020, doi: 10.1016/j.jobe.2020.101315.

[46] J. González Ordiano, S. Waczowicz, V. Hagenmeyer, and R. Mikut, “Energy

forecasting tools and services," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 2, 2018, doi: 10.1002/widm.1235.

[47] B. M. G. Mataloto, "IOT*(AMBISENSE) – Smart Environment Monitoring using Lora,", master thesis at METI in ISCTE-IUL, 2019.

[48] I. J. Lean, T. B. Farver, H. F. Troutt, M. L. Bruss, J. C. Galland, R. L. Baldwin, C. A. Holmberg and L. D. Weaver, "Time Series Cross-Correlation Analysis of Postparturient Relationships Among Serum Metabolites and Yield Variables in Holstein Cows," J. Dairy Sci., vol. 75, no. 7, pp. 1891–1900, 1992, doi: 10.3168/jds.S0022-0302(92)77949-1.

[49] S. Akhtar and H. G. H. H. Mohammad, "Time series cross-correlation analysis of HIV seropositivity and pulmonary tuberculosis among migrants entering Kuwait," *Int. J. Mycobacteriology*, vol. 1, no. 1, pp. 29–33, 2012, doi: 10.1016/j.ijmyco.2012.01.005.

[50] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Noise Reduction in Speech Processing," *Noise Reduct. speech …*, vol. 2, p. 229, 2009, doi: 10.1007/978-3-642-00296-0.

[51] M. Cools, E. Moons, and G. Wets, "Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models," *Transp. Res. Rec.*, no. 2136, pp. 57–66, 2009, doi: 10.3141/2136-07.

[52] S. F. Crone and S. Pietsch, "A naïve support vector regression benchmark for the NN3 forecasting competition," *IEEE Int. Conf. Neural Networks - Conf. Proc.*, no. September 2007, pp. 2454–2459, 2007, doi: 10.1109/IJCNN.2007.4371343.

[53] R. J. Hyndman, "Cross-validation for time series", 2016. Available: https://robjhyndman.com/hyndsight/tscv/. [Acessed: September, 15th 2020].

[54] R. Bardenet, M. Brendel, B. Kégl, and M. Sebag, "Collaborative hyperparameter tuning," *30th Int. Conf. Mach. Learn. ICML 2013*, vol. 28, no. PART 2, pp. 858–866, 2013.

[55] H. D. Acquah, "Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship.," *Dev. Agric. Econ.*, vol. 2, no. 1, pp. 001–006, 2010, doi: 10.1016/j.fishres.2005.08.011.

Appendices

A. Temperature evolution

From September 2016 to September 2019 in Lisbon in Cº.

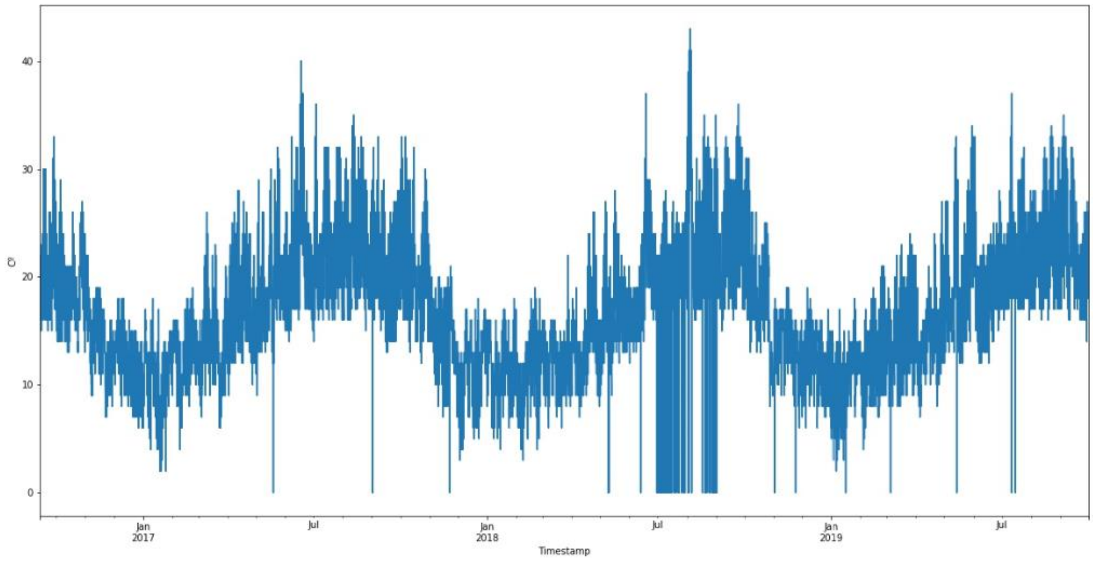

Figure 17 - Temperature evolution in Lisbon from September 2016 to September 2019

B. Humidity evolution

From September 2016 to September 2019 in Lisbon in percentage.

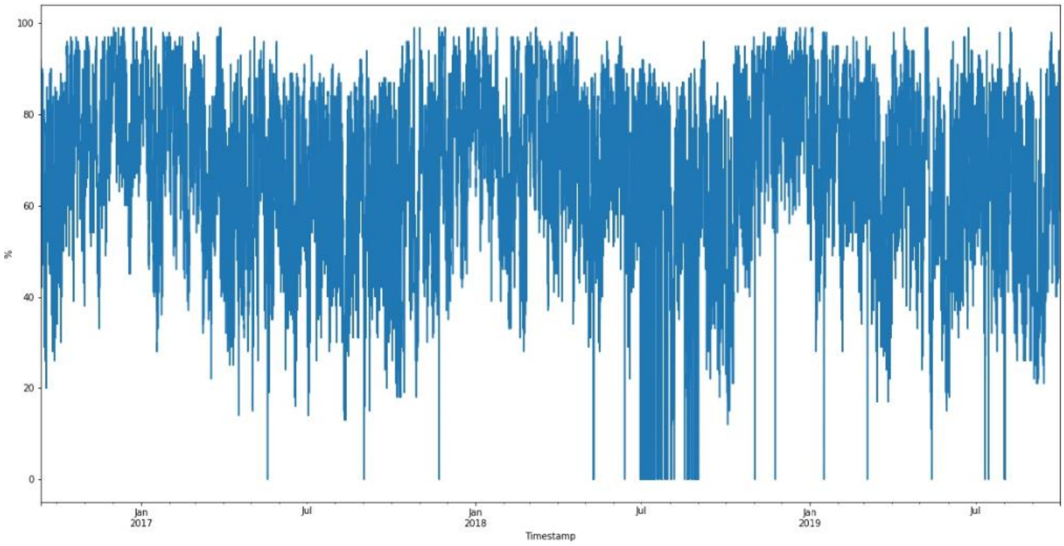

Figure 18 - Humidity evolution in Lisbon from September 2016 to September 2019

C. Precipitation evolution

From September 2016 to September 2019 in Lisbon in mm

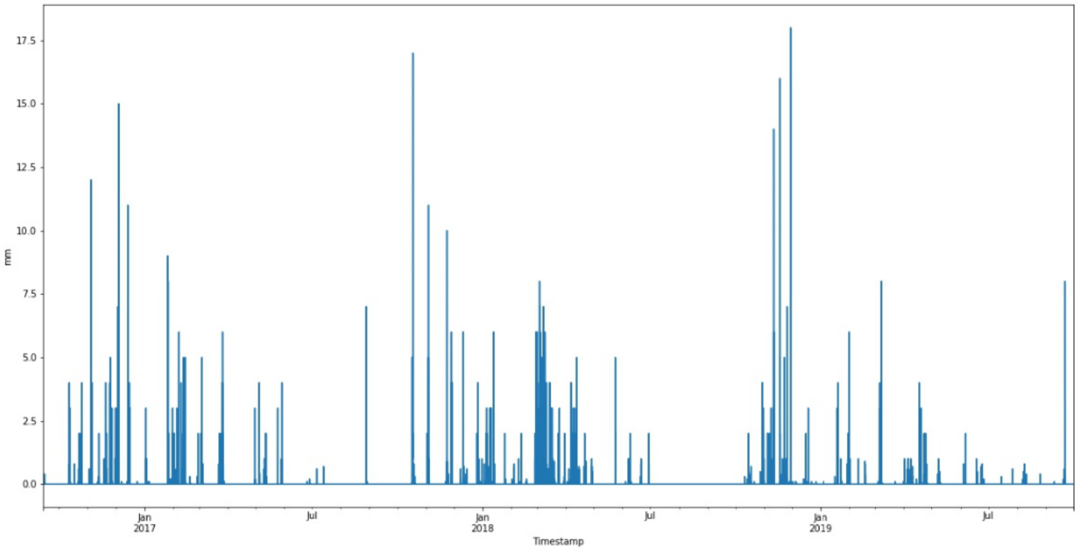

Figure 19 - Precipitation evolution in Lisbon from September 2016 to September 2019

D. kWh consumption by year

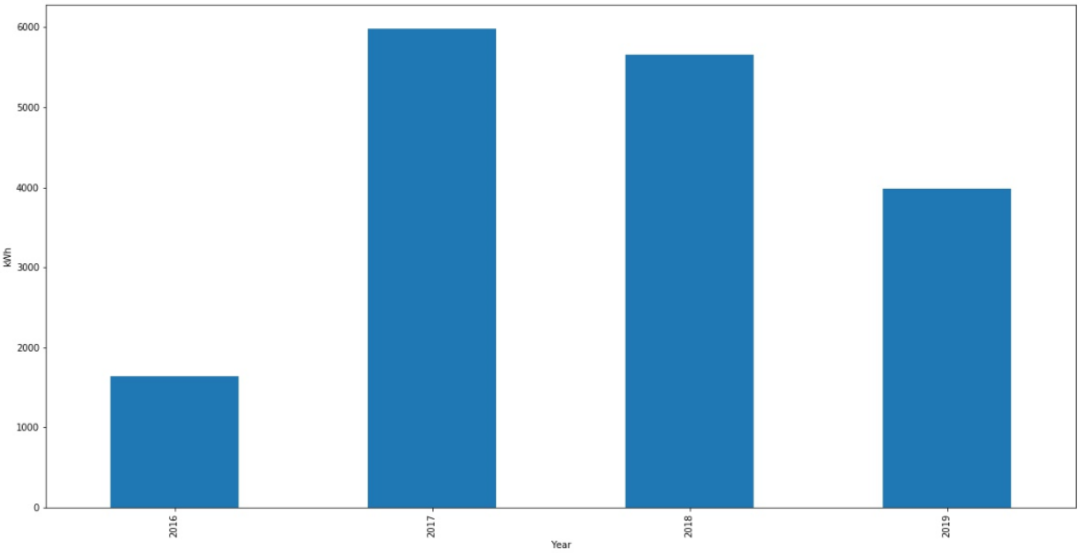

Figure 20 - Kindergarten kWh consumption by year

E. kWh consumption by month

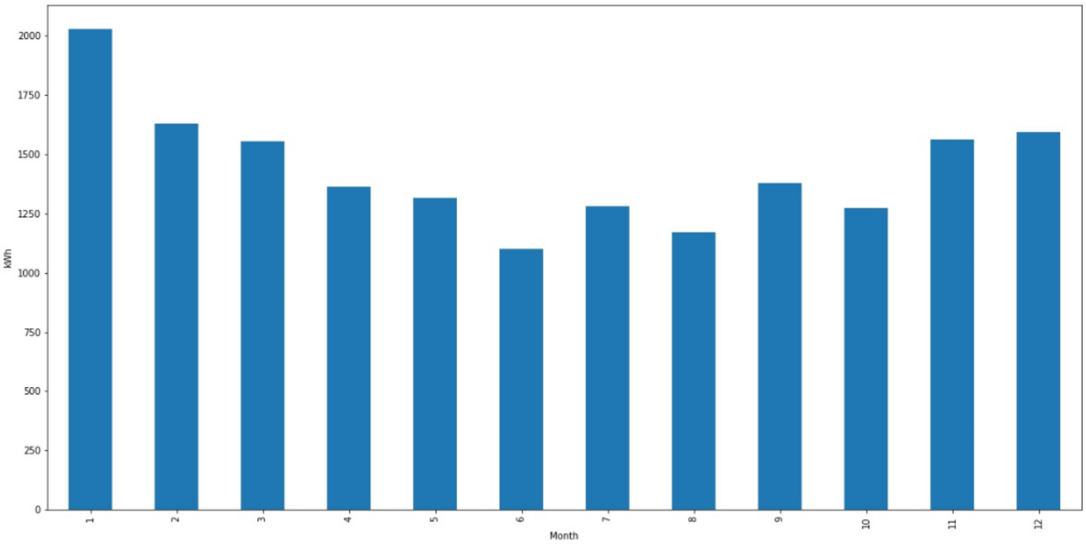

Figure 21 - Kindergarten kWh consumption by month

F. kWh consumption by day

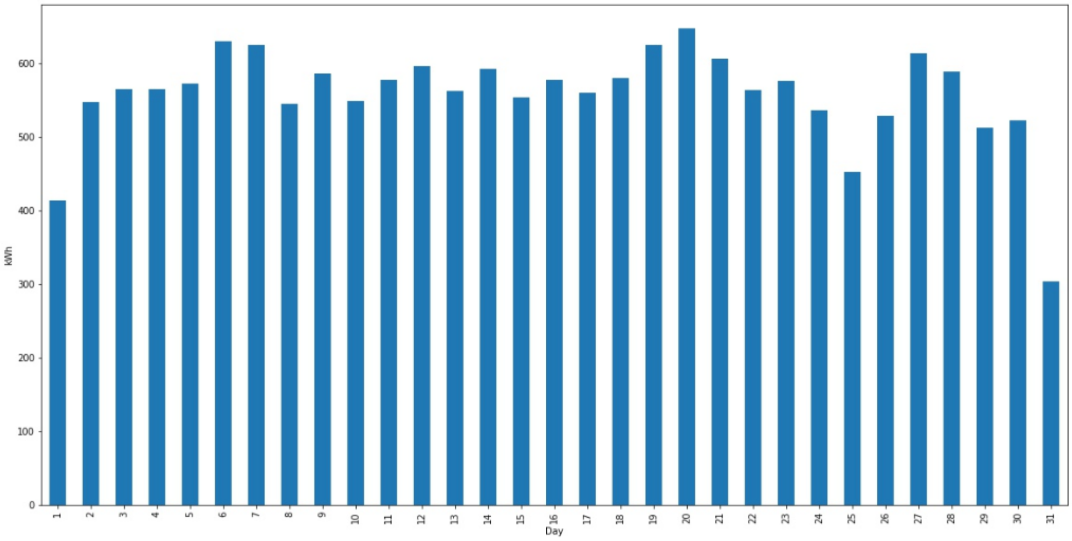

Figure 22 - Kindergarten kWh consumption by day

G. kWh consumption by hour

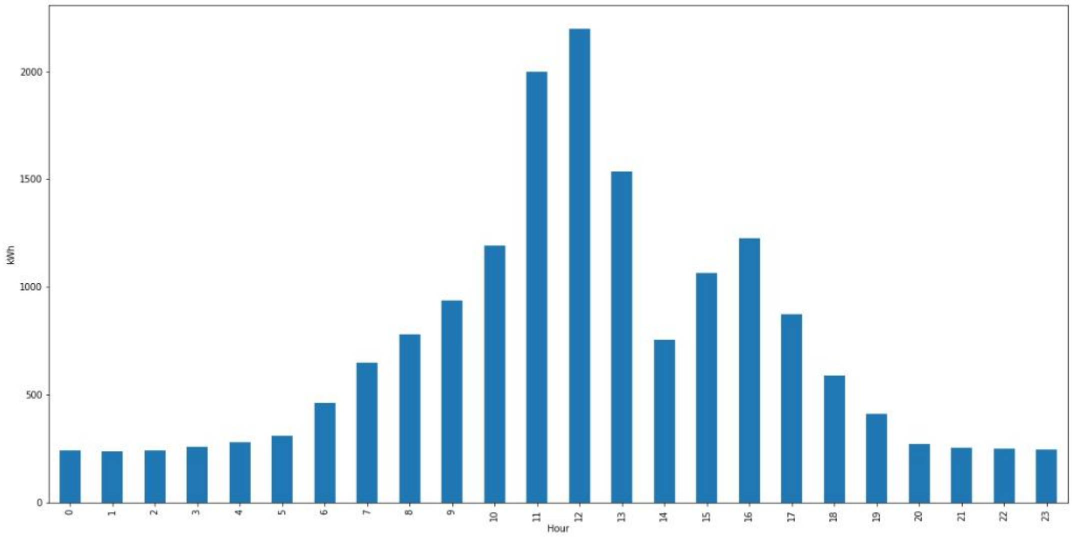

Figure 23 - Kindergarten kWh consumption by hour

H. Scatter-plot kWh and “Temperatura (Cº)”

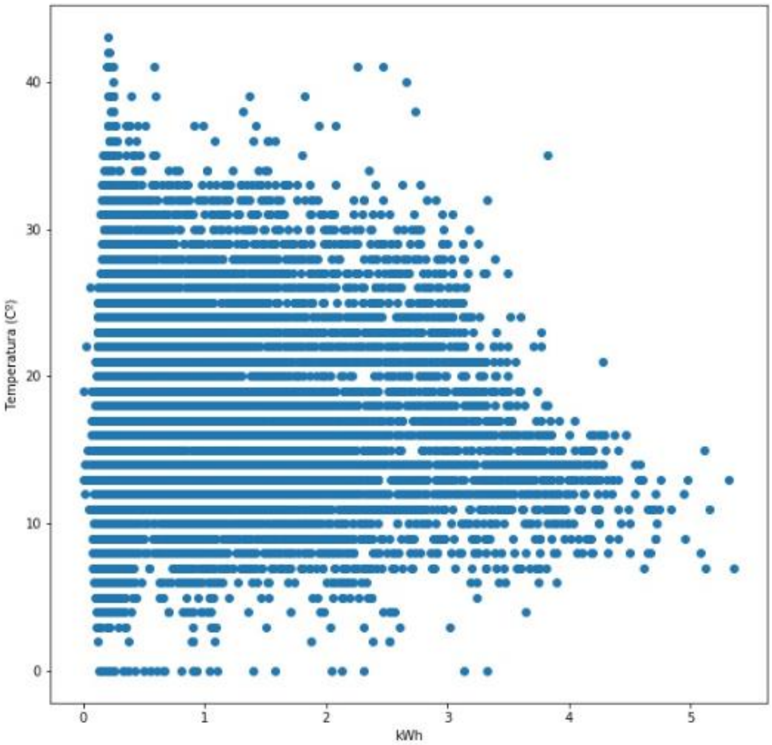

Figure 24 - Scatter-plot between kWh consumption and temperature

I. Scatter-plot kWh and “Humidade (%)”

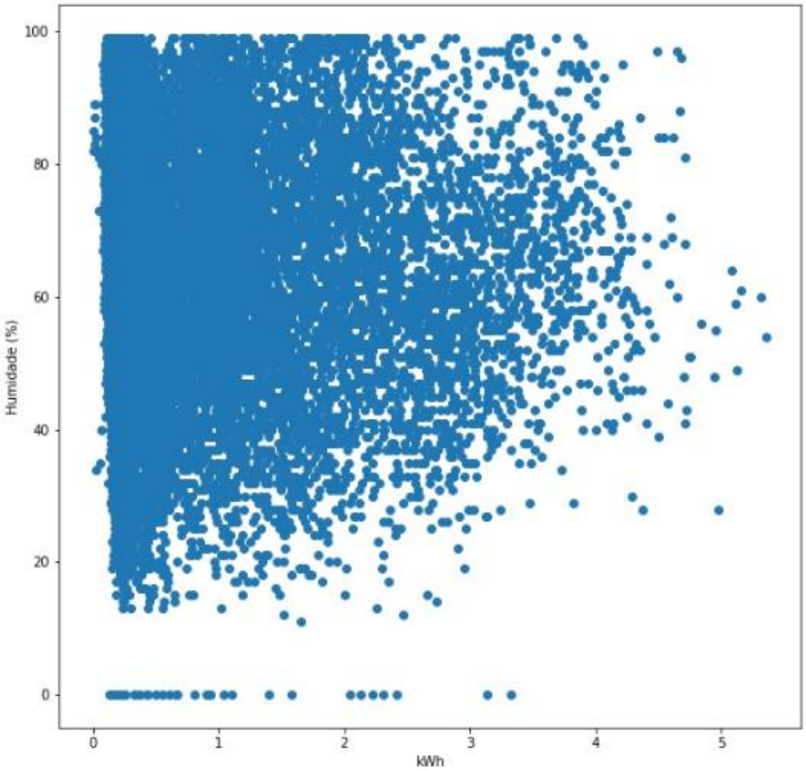

Figure 25 - Scatter-plot between kWh consumption and humidity

J. Scatter-plot kWh and “Precipitação (mm)”

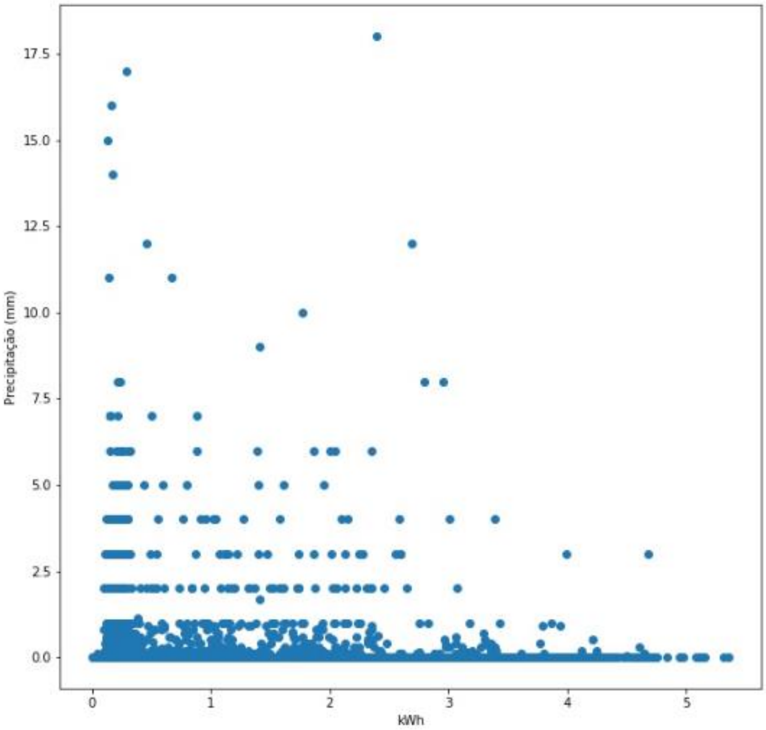

Figure 26 - Scatter-plot between kWh consumption and precipitation