Machine Learning applied to energy demand forecast in IST Alameda Campus

Francisco Delca Gouveia Pereira

Thesis to obtain the Master of Science Degree in

Mechanical Engineering

Supervisor: Prof. Carlos Augusto Santos Silva

Examination Committee

Chairperson: Prof. Edgar Caetano Fernandes

Supervisor: Prof. Carlos Augusto Santos Silva

Member of the Committee: Prof. Susana Margarida da Silva Vieira

November 2019

ACKNOWLEDGEMENTS

First of all, I would like to thank Professor Carlos Silva, for giving me the opportunity to develop this thesis, as well as, for all the advice, knowledge and motivation provided.

I would also like to thank my family for all the support. To my girlfriend and my friends that were always there for me when I needed.

ABSTRACT

Energy consumption forecasting of buildings plays a crucial role in making planning decisions by facility managers and energy providers. These decisions are used to reduce the intrinsic environmental impact of the building sector. Nowadays, with the imminent application of Building Energy Management Systems (BEMS) and the consequent increase of generated data, the use of machine learning algorithms to provide such predictions becomes a natural solution. In this study, four machine learning algorithms (MLP, SVM, RF, and XGB) were compared in three different forecasting horizons (an hour, a day, and a week) for four buildings (Civil, Central, North tower, and South tower) located at Instituto Superior Técnico, Lisbon (4 algorithms x 3 forecasting horizon x 4 buildings = 48 models). In the development of such models, three years of hourly gathered data of each building consumption and outdoor weather conditions were used. Firstly, due to the missing values presented in the data, an imputation study was carried out in order to guarantee data temporal continuity. Afterwards, based on the energy consumption analysis of each building, different features were created in attempt to describe buildings' behaviour. From the created features, different data sets were developed per building and forecast horizon, where a feature selection analysis supported with the use of a wrapper method, known as RFE, took place. With that selection, it was concluded that the most important features were the type of day, the lagged features, and the average cluster consumption of a typical working day. At last, an hyperparameter search using Bayesian optimization was conducted and the models were then used to forecast the last year of data. Among all the models used, SVM models outstood, showing higher accuracies in most of the forecasting horizons and buildings. Overall, in 93% of the forecasted days, it was achieved a MAPE error of 10.95%, 9.17%, 10.48%, and 12.66% for Civil, Central, North tower, and South tower buildings in a week horizon forecasting, respectively. In addition, it was also noticeable an increasing annual error tendency when the models attempt to predict in greater horizons.

Key-words: energy, building consumption, forecast, machine learning

RESUMO

A previsão do consumo energético dos edifícios representa uma ferramenta essencial para o planeamento e adopção de diferentes estratégias energéticas por parte dos gestores e fornecedores de energia dos edifícios, a fim de reduzir o impacto ambiental existente no sector. Actualmente, a aplicação de sistemas de gestão energética em edifícios e o consequente aumento da quantidade de dados gerados, leva ao uso de modelos orientados por dados, maioritariamente de modelos de machine learning, para obter a previsão do seu consumo. Neste estudo, quatro algoritmos de machine learning (MLP, SVM, RF e XGB) foram utilizados e comparados em três horizontes temporais diferentes (uma hora, um dia e uma semana) na previsão do consumo energético de quadro edifícios (Civil, Central, Torre sul e Torre norte) localizados no campus da Alameda do Instituto Superior Técnico, Lisboa, (4 algoritmos x 3 horizontes de previsão x 4 edifícios = 48 modelos). Para o desenvolvimento de cada modelo, três anos de dados horários das condições atmosféricas e do consumo de cada edifício foram disponibilizados. Primeiramente, devido à falha de valores apresentada nos dados, foi realizado um estudo de imputação com o objectivo de garantir a continuidade temporal. De seguida, com base na análise do consumo energético de cada edifício foram criadas diferentes variáveis com a finalidade de descrever o comportamento de cada edifício. A partir das variáveis criadas foram desenvolvidos diferentes data sets por edifício e horizonte de previsão, onde foi posteriormente utilizado um método recursivo de eliminação de variáveis como apoio à selecção das variáveis mais importantes de cada data set. Desta selecção três variáveis foram consideradas indispensáveis para a realização da previsão, tais como: o tipo de dia, o consumo da hora, do dia e da semana anterior à altura da previsão; e ainda a média diária de um conjunto de dias típicos de trabalho. Ultimamente, foi aplicado um algoritmo de optimização bayesiana para a selecção dos hiperparâmetros de cada modelo e a previsão foi então realizada para o último ano disponibilizado. Entre todos os modelos utilizados, o modelo SVM destacou-se, apresentando os melhores resultados na maioria dos edifícios e horizontes de previsão. Num âmbito geral, em 93% dos dias previstos, os edifícios de Civil, de Central, da Torre Norte e da Torre Sul, obtiveram erros médios absolutos percentuais de 10.95%, 9.17%, 10.48% e 12.66% para um horizonte temporal de uma semana, respectivamente. Foi também observada uma tendência crescente do erro anual para previsões com um maior horizonte temporal.

Palavras-chave: energia, consumo de edifícios, previsão, machine learning

Table of Contents

List of Tables

List of Figures

List of Acronyms

ANN	Artificial Neural Network
AR	Autoregressive
ARIMA	Autoregressive, Integrated and Moving Average
ASHRAE	American Society of Heating, Refrigerating, and Air Conditioning Engineers
BEMS	Building Energy Management System
CBR	Case-based reasoning
CV	Cross-Validation
CV(RMSE)	Coefficient of Variation of the Root Mean Squared Error
DE	Differential Evolution
DT	Decision Tree
ECD	Energy Consumption Dataset
GA	Genetic Algorithm
HMM	Hourly Monthly Mean
HVAC	Heating and Ventilation Air Conditioning
iPSO	improvement Particle Swarm Optimization
LS-SVM	Least Square Support Vector Machine
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MF	Miss Forest
MICE	Multiple Imputation by Chained Equations
ML	Machine Learning
MLP	Multiple Layer Perceptron
MLR	Multiple Linear Regression
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
RF	Random Forest
RFE	Recursive Feature Elimination
SA	Simulated Annealing
SVM	Support Vector Machine
TLBO	Teaching Learning Based Optimization
ts-CV	time series - Cross-Validation
t-SNE	t-distributed Stochastic Neighbor Embedding
WCD	Weather Conditions Dataset
XGB	Extreme Gradient Boosting

Chapter 1

Introduction

1.1 Motivation

Over the last decades, technological and social advancements have enlarged humankind capability to have better living conditions, which lead to an increase of overall average human lifespan and resulted in a current world population of over than 7 billion, expected to achieve around 10 billion by 2050 [1].

Due to this rapidly population growth and people's tendency to move to urbanized areas, the city's ability to full-fill every inhabitant's needs, is considered to be one of the twenty-first century biggest challenges [2]. Cities are responsible for providing services in terms of, transportation, healthcare, safety, education, water supply, and most importantly, energy. According to European Commission, nowadays in Europe, cities accommodate around 75% of the population, being liable for about 70% of the global energy consumption and CO_2 emissions [3].

This large energy consumption can be divided in three main sectors: industrial, transportation, and buildings [4]. Nowadays, the buildings sector, as reported in Figure 1.1, accounts with almost 40% of energy consumption and 36% of CO_2 emissions [5].

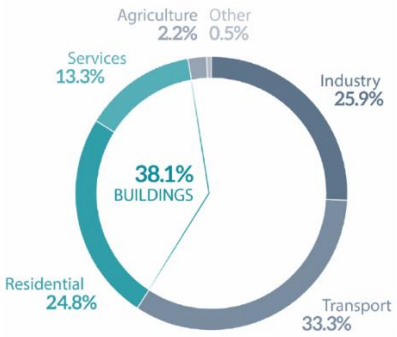


Figure 1.1 - Energy consumption, percentage per sector – EU-28 – 2014 [6]

The buildings sector can be split into two different types of buildings, residential and non-residential. The latter group, often referred to as the services sector consists of businesses institutions and organizations that are distinguished by the services that provide, such as hospitals, universities, hotels, and restaurants. The forms of energy that are predominant in this sector are electricity and natural gas (used for space heating, water heating, lighting, cooking, and cooling) accounting with 10 to 15% of the overall energy consumption [6] and growing by an average of 1.8% per year from 2010 to 2040 [4].

It is imperative to state the need to reduce building excessive energy consumption, since they represent a significant fraction of the overall energy expenditure, which consequently results in high environmental impacts.

Since 2006 [7], the European Commission has been implementing energy efficiency measures for sustainable development, being one of the main objectives to reduce the annual energy consumption by 27% till 2030.

In order to achieve that goal, in the building sector, it is essential to know, that in Europe, not only 35% of the buildings are over 50 years old, but around 75% are energy inefficient [5]. One way to improve energy performance of existing buildings is to effectively predict its consumption, enabling the endorsement of diverse operating strategies to increase energy efficiency and to detect faults related to systems malfunction.

Buildings energy behaviour is influenced by several factors, such as atmospheric conditions, building construction and consequent thermal properties, occupancy, lighting and other systems, in particular, heating, ventilating and air conditioning (HVAC) systems. These last two factors are responsible for approximately 50% of the services sector electricity expenditure [8]. As a result of this wide range of influenceable factors, the building system becomes inevitably complex, hampering the task of performing a fast and accurate consumption forecasting.

To address this need over the past 50 years [9], a large number of investigations have been carried out to ascertain the complexity related with buildings energy consumption and to find out an accurate representation of its energy performance. Currently, building energy simulation can be branched into three different approaches: **white box**, **grey box** and **black box**. The approaches names are related to the insight that the user has about what the model is doing, being the **white box** the clearest one. In Table 1.1 a summary of the comparison between the different approaches mentioned is provided.

Table 1.1 - Summary of comparison of different approaches for the analysis of buildings' energy consumption [10]

Type of Approach		Input data	Model and Software	Execution	Computational Cost	Accuracy
White box	Detailed Simulation	Detailed Physics Information	DOE-2, EnergyPlus, TRYSYS, ESP-r	Hard	High	High
	Simplified Simulation		Degree day method, Temperature frequency method, Residential load factor method	Easy	Low	Reasonable
Grey box		Physics Information and Historical Data	RC network	Hard	High	Reasonable
Black box		Historical Data	ANNs, SVM, statistics regression (ARIMA), cluster	Hard	Low, except SVM	High, expect regression models

White box approaches, also known as physical models, are widely used in engineering and are grounded by thermodynamic laws, requiring many building details and surrounding environmental conditions as input data. Computationally they are very expensive, and data input requirements may,

in some circumstances, not be entirely fulfilled. Lately, these approaches have been simplified to reduce the computational cost, although they are error-prone and usually overestimate the energy expenditure of buildings [11]–[13].

Grey box approaches merge the models mentioned above with statistical modelling, allowing the use of simplified building information and historical data to perform the energy simulation. Nevertheless, they provide reasonable accuracy predictions with high computational cost depending on building information [14]–[16].

In order to circumvent the shortcomings referred by the first two approaches, **black box** approach was employed. This purely data-driven approach, when compared with the others, is able to develop a faster and higher accurate consumption forecasting, based only on historical data, avoiding thus the need of physical building details [17]. For those reasons the models that characterize this approach, mainly in machine learning field, have been receiving particular attention in the past years.

1.2 Objectives

This study aims to develop and compare four machine learning models (ANN(MLP), SVM, RF, and XGB) in three different forecasting horizons (an hour, a day, and a week) for four buildings (Civil, Central, North tower, and South tower) located at Alameda campus of Instituto Superior Técnico, Lisbon. In the development of such models, three years of hourly gathered data of each building consumption and outdoor weather conditions was used. From the three years, the first two (2014 and 2017) were used for the training stage and to perform different strategies in order to enhance each of the models, and the last year (2018) was used to test the capabilities of each of the developed models.

1.3 Contributions

From this study, four contributions may be stood out:

- An imputation study for this study dataset was developed.
- Adaptation for the first time of XGB model to forecast building energy consumption in hourly granularity.
- Development of forecasting models in three different horizons for four buildings located at Alameda campus, IST, Lisbon.
- The code developed along this study is publicly available in [18].

1.4 Structure of the thesis

This work was organized in five main chapters:

- **State of the Art** - A literature review of studies done prior to this work, their achievements and conclusions. It was also supplied a brief explanation about the machine learning algorithms used, as well as, the complementary models and the evaluation error metrics.

- **Study Case** - This chapter gives an introduction and an energy analysis of the different buildings used to test the chosen machine learning algorithms.
- **Methodology** - It intends to clear and explain every strategy employed to achieve the objectives of this work.
- **Results** - This chapter presents a detailed analysis about the results obtained in all the strategies adopted. The final forecasted results per building and time horizon were also displayed and commented.
- **Conclusions** - This final chapter gives a reflection about the results by understanding each models' prediction and limitations. Future recommendations were also given to about any future work.

Chapter 2

State of the Art

2.1 Concept of Intelligent Energy Management

The disruptive technologies of the twenty-first century have forced the replacement of conventional power grids, that are ill-suited for the needs of today's electricity sector, to smart grids. The core of this technology allows bidirectional communication that has the potential to optimize the linkages between energy supply and demand through digital information [19]. Among the technologies that can benefit from smart grids capabilities, Building Energy Management System (BEMS) is a primary candidate since a vast majority of the potential customers are buildings.

BEMS emerged due to the increased awareness of the environmental impact of energy use and generation in the sector. This system combines hardware and software solutions to improve energy management by converging the data gathered from existing sensors throughout the building into a centralized control. This centralized control is then responsible for the real-time monitoring of the diverse existent components of the building, providing detailed reports and recommendations for further energy and cost improvements without compromising efficiency and comfort of its users.

However, this type of systems contribute to better energy use polities, their performance frequently lacks the expectations, mostly due to the incapability of finding a suboptimal operation spot when handling with extensive amounts of continuously changing data, caused by the dynamic and uncertain of indoor and outdoor conditions [20].

To overcome the vast amount of generated data, the use of data-driven models to support BEMS seems like a natural solution. A recent study stated [21], that the inclusion of those models in BEMS would allow up to 22% of energy savings in the European building sector by 2028. This might be explained by their capability of providing diverse tools, such as anomaly detection and consumption forecasting, to strength decision making towards reducing energy expenditure. A framework of BEMS supported by data-driven models is shown in Figure 2.1.

This dissertation goal is to study the applicability of data-driven models, in one of the diverse tools that enhance energy use, specifically building energy consumption forecasting.

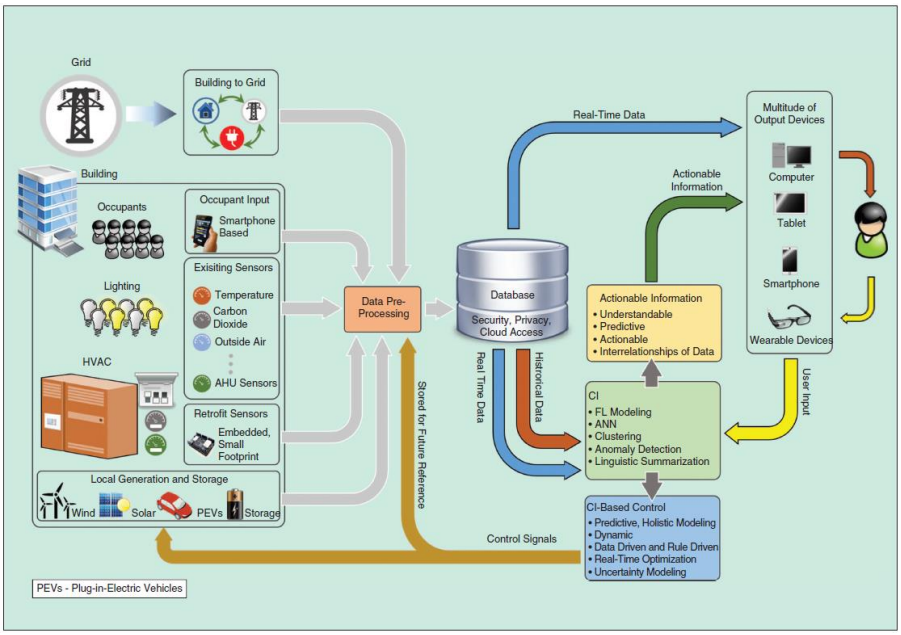


Figure 2.1 - Framework of BEMS supported by data-driven models [20]

2.2 Forecasting Data-Driven Models

As it is known, data-driven models, instead of using detailed building information to develop an energy analysis, use only historical and available data to learn the dynamic energy behaviour of the buildings and are often referred to as empirical models. Nowadays, due to their ability to extract useful information at low cost, they have been applied in diverse fields such as commerce [22], political campaigns [23], and medical diagnosis [24].

The most common data-driven models used for energy consumption forecasting may be ramified into two fields: the statistical field and the machine learning field. From the statistical field, the models often applied were the autoregressive, integrated and moving average (ARIMA) and the multiple linear regression (MLR). On the other hand, from the machine learning field, two models were substantially applied, specifically, artificial neural networks (ANN) and support vector machines (SVM), and another one least used named as ensemble model, Figure 2.2.

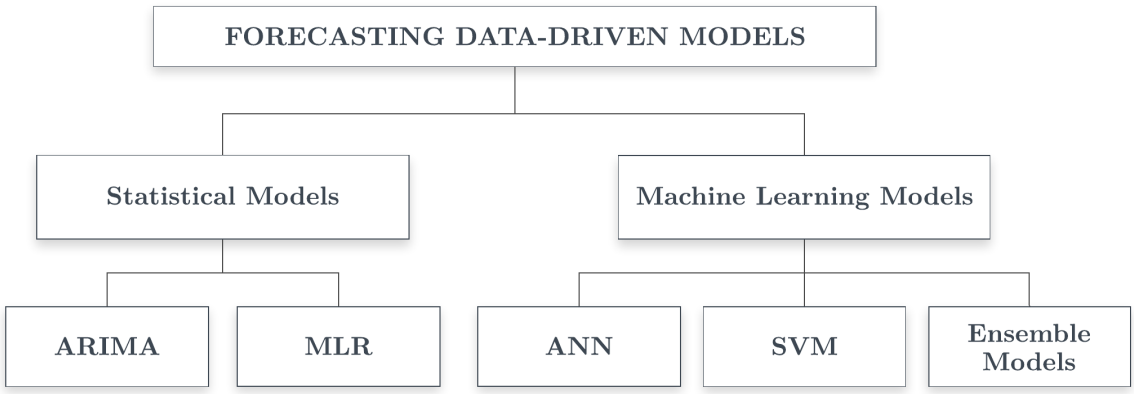


Figure 2.2 - Summary of data-driven models used to predict building energy consumption

These models methodology may be briefly divided into four main sequential steps:

1. **Data acquisition** - performed by sensors and meters, this step is responsible for gathering the necessary data to feed the models (input and output data). In most case scenarios, the accuracy of the models increases when the data is well collected and related with the nature of the desired output [25];
2. **Data pre-processing** - includes all the measures applied to the acquired data that aimed the quality, consistency and appropriate format to feed in the model. It may include processes such as erase the presence of noisy and missing data, interpolation, and normalization;
3. **Model training** - is a requirement in the development of data-driven models and its where the model learns and estimates parameters from the input data to enhance the prediction;
4. **Model testing** - is when the model predictions are compared with the output data intended for testing, this comparison is made through standard error metrics.

Furthermore, it is necessary to specify the temporal granularity of the forecast, which not only influences the input data and data pre-processing, but also the error tolerance and accuracy of models. It can be divided into three groups [25]:

- **Long term**: from one year to ten years forecast. Typically used for planning new infrastructures;
- **Medium term**: from monthly to annual forecast, mostly used for an efficient operation planning and maintenance of energy systems;
- **Short term**: hourly, daily or weekly forecast. Frequently used to adjust the energy demand and supply to the grid, as well as to ensure the expected performance of the different power systems.

In this dissertation, the energy consumption forecasting is framed in the short-term granularity, therefore all the revised literature focuses only on that granularity.

2.2.1 Statistical Models

The statistical models that have been frequently used to predict building energy consumption are generally regression models [26]. Statistical regression techniques find relationships between the different variables through mathematical formulations to predict a specific target. Several investigations, took advantage of this approach to address diverse challenges in the analysis of building energy behaviour, for example, to predict energy used through simplified variables, foresee building energy index, and estimate significant energy parameters for analysis [9].

From regression models, there are at least two models that are mandatory to emphasize, the MLR and the ARIMA. The latter was specifically created to handle and correlate time series data for prediction. Examples of its applicability in short term building energy prediction may be found in [27]–[30].

Although these models are easy to develop and use, they lack on flexibility in coping with the nonlinearity often found in building energy consumption. In consequence of that, the statistical approach presents poorer prediction accuracy, which limits its applicability, when compared with machine learning models. For example, in 2011, *Penya et al.* [31], compared two statistical models, ARIMA and autoregressive (AR), with an ANN, to forecast the consumption of an institutional building, in Bilbao. The machine learning model reached higher accuracy values than both statistical models. The same conclusion was achieved, in a university in Girona, 2015, by *Massana et al.* [32] when comparing an MLR

with two machine learning models (ANN and SVM). Moreover, in [33], the same machine learning models outperformed ARIMA in the energy prediction of an office building.

Nevertheless, in 2016, an ANN and a statistical approach was used to predict the energy consumption of two of the buildings tested in this work. The study concluded an overall improvement of about 10% when using ANN instead of the statistical model [34].

Consequently, as it is noticeable from the studies reviewed in this section, there is a visible outperformance of the machine learning over the statistical approaches, with that in mind, the following section will be focus on machine learning models.

2.2.2 Machine Learning Models

Machine learning is an interdisciplinary field based on statistics and optimized mathematics techniques which gives computer systems the ability to learn and improve performance on a given task, being only fed with data without the need to be explicitly programmed [35].

This field may be divided into two categories:

Supervised Learning is used when the input data that feeds the model has the corresponding solution, output data. The objective of this category is to learn a mapping function (f) from the input (X) to the output data (Y), that is $Y = f(X)$. This type of learning can be organized according to the desired output:

- Regression: the output data is quantitative, consisting of real values, which may be integers or floating points. An example of that can be found in this study primary objective, where the desired output is the buildings' energy consumption predicted values;
- Classification: unlike regression, the output data is qualitative, consisting of discrete or categorial variables, that are pre-established by the user. For example, a model which classifies a day type as weekday or weekend based on is daily consumption.

Unsupervised Learning is applied when only the input data is available, and the goal is to reveal hidden structures or patterns in data. It is branched into:

- Clustering: the data is divided into clusters, e.g. k-means;
- Density estimation: aims to estimate the distribution of the data in some space;
- Dimensionality reduction: maps the data into a lower-dimensional space, frequently used for simplification, e.g. principal component analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE).

There is also another category, known as reinforcement learning, that will not be discussed once is outside the scope of this project.

Therefore, as mention earlier, being our primary objective the prediction of energy consumption, the main machine learning models presented in the following section are within the supervised regression learning category.

Artificial Neural Networks (ANN)

Artificial neural networks model is a non-linear supervised learning algorithm inspired by the biological neural network that constitutes animal brains. Analogous to biological neurons, ANN uses interconnected artificial neurons - called processing units - that are grouped in layers. Where, each connection is associated with a numerical value designated by weight, each processing unit has an activation function, and by mathematical convention, each layer has a bias term that stores the value of +1. Typically, three distinguished parameters are used to define a neural network: the architecture, the learning process of weights updating, and the activation function which converts the weighted processing units input into its output. Figure 2.3 shows a schematic diagram of an ANN, also known as multilayer perceptron (MLP), with one hidden layer.

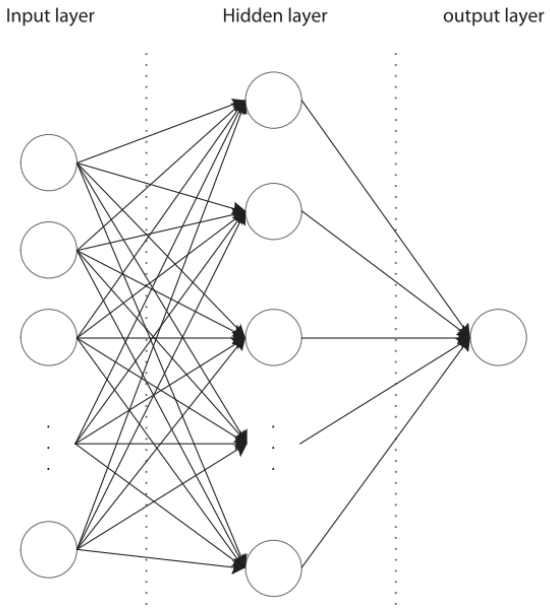


Figure 2.3 - Schematic diagram of one hidden layer ANN architecture [36]

Within artificial neural networks universe, there are several architectural typologies that have been developed to overcome diverse type of problems. Nowadays, it may be ramified into recurrent and feedforward neural networks. In recurrent neural networks, the connections can be made between all neurons, while in feedforward neural networks the connections propagate unidirectionally from layer to layer, preventing that in the same layer the output of one neuron does not influence the output of another neuron.

Within machine learning models, ANNs have been particularly popular and applied to forecast buildings energy consumption [26]. In 2005, *Gonzales* and *Zamarreno* [37], used a feedback ANN model developed in [38] to predict next hour consumption of an institutional building, concluding that increasing the number neurons is not directly proportional to better predictions.

In the following year, *Karatasou et al.* [39] study the applicability of statistical techniques, such as hypothesis tests, information criteria, and cross-validation, described in [40], to distinguish the most relevant features and useful hidden layers. The ANN model improved is accuracy with this procedure for a day and an hour horizon forecasting.

Then in 2009, *Yokoyama* [41] used an optimization method known as modal trimming [42], instead of the typical gradient descendent in ANN backpropagation stage, to predict the cooling load of a service building. The method used yield better results than the typical ANN.

Furthermore, in 2011, *Guillermo et al.* [43], predicted a day a-head energy consumption of an institutional building, in Valencia. To predict the total consumption forecast it was used a distinct ANN for each of buildings end-use energy and an adequate selection of the training data set to simplify the ANN architecture. This selection had the goal to elect the most similar days to the one to forecast through two parameters: the labor activity parameter that characterizes the occupancy patterns and the temperature coefficient to define the weather conditions. The use of those parameters yielded better predictions over the usage of the entire data set. Seeking for better results, in [44], was introduced a temperature curve model to enhance the training data set selection, nevertheless no significant improvements were achieved.

After that, in 2014, *Mena et al.* [45], applied one hidden layer ANN model to forecast an hour a-head electric consumption of a bioclimatic building, in Spain. For the input data selection, two free model analytical tools were used: correlation and mutual information. The application of these tools yield better results than the usage of the all the available input data.

In 2015, at least two researches concerning the short term energy consumption forecasting using ANN were disclosed. Firstly, by *Platon et al.* [46], in the comparison of an ANN with a case-based reasoning algorithm (CBR) for 1 to 6 hours a-head prediction of an institutional building. In this study the ANN stood out with better accuracy values. Additionally, in order to reduce the computational cost associated with the vast amount of input data, a principal component analysis (PCA) was successfully implemented without compromising the forecast accuracy. After that, *Li et al.* [47], appealed for the application of evolutionary algorithms to data-driven models by using an ANN with an improved particle swarm optimization (iPSO) instead of the default gradient based method. The optimized model was more effective than the traditional ANN algorithm in the prediction of an hour a-head consumption of two institutional buildings. From this study, it was also concluded that for online predictions, the iPSO is more suitable tan the genetic algorithm (GA), due to its rapid convergence on searching the optimal solution.

In the following year, *Chae et al.* [48], when forecasting a day a-head energy consumption of a service building, tested the applicability of an ANN with Bayesian optimization to improve model generalization. To input relevant data to the model it was used a feature extraction technique by means of an ensemble machine learning algorithm, known as random forest (RF). The study revealed a decreased in the forecast error as the number of weeks of data available for training increased.

Lately, in 2018, *Li et al.* [49] used another evolutionary algorithm called teaching learning based optimization (TLBO) to improve an ANN learning process. The model was compared with the previously used methods in [47] for the same buildings, and presented superior results, in terms of computational speed and accuracy.

Support Vector Machines (SVM)

The SVM model is a supervised machine learning technique, well known for its robustness and accuracy, even when the data required for the learning process is reduced and arbitrarily structured. Due to its ability to solve nonlinear classification and regression problems through the use of a kernel

function, its application has become increasingly common in both research and industry fields since it was first implemented in 1995 [50].

The main objective of the regression SVM model is to determine the function whose measured output data deviation does not exceed the error term (predefined) for each of the input data combinations. In the context of nonlinear regressions, data is transformed by a function called kernel, which maps input data into a higher dimensionality space, where linear regression is later applied. Although this mapping results in a lack of transparency of what is going on within the model, it deals with the non-linearity often encountered in complex problems. There are three types of kernel functions: linear, polynomial, and radial. The main challenge of this algorithm is the selection of kernel function and its parameters according to the nature of the problem since they greatly affect is performance. For this reason, optimization techniques such as evolutionary algorithms are often used [51]. In the following paragraphs, some application examples of SVM in building consumption forecasting are shown.

Li et al. [52], in 2009, conducted a study comparing the traditional ANN with an SVM model, to predict an hour a-head cooling load of an office building, in China. The SVM model used radial basis function kernel and revealed higher accuracy than the traditional ANN. In the same year, the same type of conclusion was achieved when performed in another office building by *Xuemei et al.* [53]. The research compared another strand of SVM model, denominated by least square support vector machine (LS-SVM), with a traditional ANN. The new model showed better generalization performance and accuracy in four different error metrics. In 2010, *Li et al.* [54] improved is previous work [52] by using a combination of two different algorithms: simulated annealing (SA) and PSO to select SVM parameters.

Furthermore, in 2015, *Fu et al.* [55] presented a study using an SVM to predict the next day electrical consumption for different public buildings. Four months of data were used. SVM model results have transcendent in accuracy over ARIMA, decision trees (DT) and ANN. That same year, *Massana et al.* [32] tested diverse models, such as MLR, MLP and SVM to predict an hour a-head energy consumption. SVM model outstood among the others, presenting the best trade-off accuracy/computational cost only using the actual building occupancy and the outside temperature, as input variables. It was also possible to conclude that the use of building interior conditions as input did not improve the model performance when HVAC operation conditions were predefined.

Ensemble Models

Ensemble models, as a more advanced technique, was introduced in 1990 [56]. In machine learning, is defined as an approach that uses multiple learning algorithms to obtain a better accuracy performance than that could be obtained from any of the constituent learning algorithms [57]. In addition, in a regression scenario, the objective function of the combined models is to minimize the overall prediction error of the ensemble model. With that in mind, to each of the constituent models is assign a weight based on its accuracy. Hence, the one with the highest weight presents the least prediction error.

Based on the selection of the models, ensemble models can be ramified into two groups [58]:

- **Homogeneous** - is characterized by using the same learning algorithm on different subsets of the training set;
- **Heterogeneous** - uses diverse learning algorithms that are trained with the same data set.

Both groups have been used to forecast the building energy consumption, some examples of their application for short term prediction are mentioned below.

Fan et al. [51], used a homogeneous model to predict half-hourly a-head energy consumption of an institutional building, in Singapore. The forecasting ensemble model used was a weighted SVM model with nu-SVM and epsilon-SVM. It was also done a comparison between three different evolutionary algorithms, such as GA, PSO and differential evolution (DE) to determine the weights of each SVM model. The evolutionary algorithm that best suited the problem was DE. On the other way, *Xiao et al.* [59] used a heterogenous ensemble model by combining eight different predictive models, to forecast a day a-head building energy consumption. The weight of each model in the final prediction where optimize using a GA. The research concludes that the accuracy of the ensemble model is evidently better than any of the individual models.

Furthermore, in the **homogeneous universe**, the ensemble models may use two types of learning procedure which characterizes the order that each model is trained, namely, bagging and boosting. In bagging each model learns with a random subset of training data in a parallel way, e.g. random forest (RF). On the other hand, in boosting each model learns from mistakes made by the previous model in a sequential way, e.g. extreme gradient boosting (XGB).

Both random forest and extreme gradient boosting algorithms have been recently used in building energy consumption prediction [60]–[62]. One of the reasons that encourage their application was the previously used of their based model, named as decision tree (DT), in the field [26].

Decision Tree is a supervised machine learning algorithm that builds a regression model in the form of a tree structure. Basically, it breaks down the data set into smaller and smaller subsets while at the same time an associated tree is incrementally growing. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, and a leaf node represents a decision. The topmost decision node in a tree corresponds to the best predictor named as root node. A schematic diagram of a decision tree is shown in Figure 2.4.

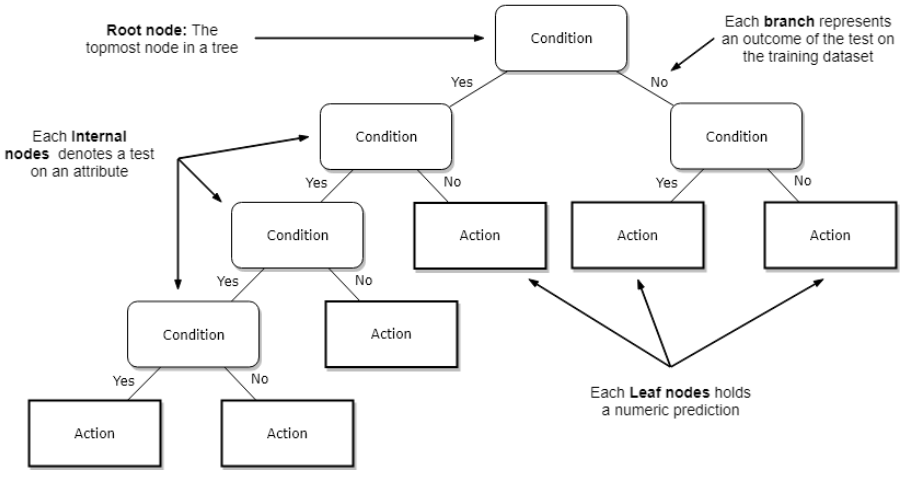


Figure 2.4 - Schematic diagram of a decision tree

Random Forest is an ensemble machine learning algorithm that uses bagging in many individual decision trees and afterwards with each decision tree prediction, its selected a final candidate based on a majority vote, Figure 2.5.

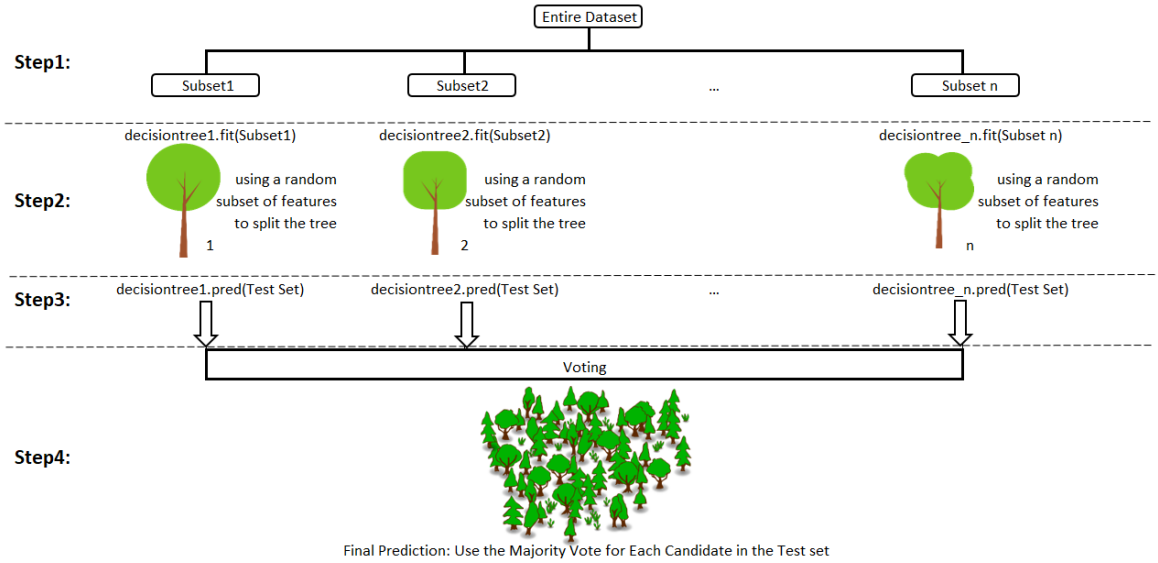


Figure 2.5 - Schematic diagram of Random Forest [63]

The application of this ensemble method can be found in [60], where *Ahmad et al.*, compared the performance of the traditional ANN with RF, for predicting the hourly HVAC energy consumption of a hotel in Madrid, Spain. As a result, ANN performed slightly better than RF. However, the ease of tuning and modelling in ensemble models stood out against the ANN, since both models had comparable predictive accuracies.

Additionally, *Wang et al.* [61] used RF, regression tree (RT), and SVM, to forecast the hourly electricity usage of two institutional buildings, in USA. The prediction performances of RF, measured by a performance index, were 14-25% and 5-5.5% better than RF and SVM, respectively. Apart from RF outperformance, it was also concluded that the most influential input data varies depending on the semester where the prediction was performed.

RF prediction performed, measured by a performance index, were 14-25% and 5-5.5% better than RT and SVM, respectively. Apart from RF outperformance, it was also concluded that the most influential input data vary depending on the semester where the prediction is performed.

Extreme Gradient Boosting is an ensemble machine learning algorithm, that uses boosting with individual decision trees. With boosting the algorithm learns from previous mistakes, using the residual error directly, as illustrated in Figure 2.6. Afterwards, the prediction is made by simply adding up all tree's predictions.

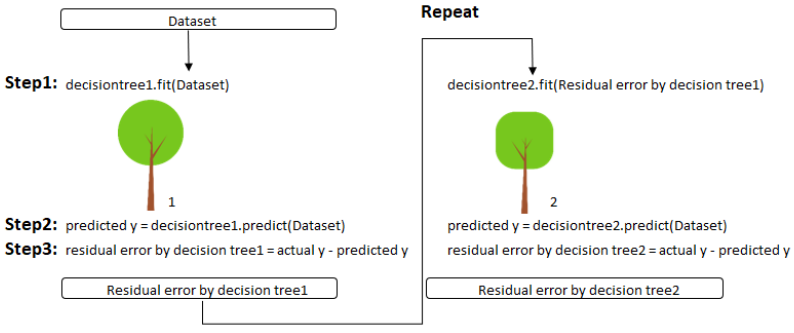


Figure 2.6 - Schematic diagram of Extreme Gradient Boosting [63]

This algorithm has been only used recently in this field, by *Robinson et al.* [62], to predict the distribution of energy intensities in cities, based on existing service buildings energy consumption. However, this study is not directly enclosed in this dissertation goal, it shows that XGB outperformed four other machine learning algorithms, such as RF, linear regression, and ANN.

Summary

In the previous sections, a detailed review of the latest machine learning algorithms application in short term building energy forecast was made. In this section, an overview of each model performance and input data selection will be considered.

In general, ANN provide accurate results in the presence of complex nonlinear problems with acceptable computational time, as can be seen in [37], [39], [43], [45]–[49], where MAPE results, do not exceed 8.38% and 13.39% for one hour and a day ahead forecast, respectively. Similarly, SVM achieved great prediction results, when applied individually such as, in [55] and [32], with MAPE values around 0.06% and 15.2% for one hour and a day a-head forecast, respectively, it is mandatory to recall that in [55] the real hourly occupancy rate was one of the input variables, which may be the reason of such good results.

In terms of ensemble models, they have provided interesting results in [51], [59]–[61], obtaining MAPE values below, 4%, 7.75%, and 4.6%, for half-hour, one hour, and a day a-head forecast. Although it is important to mention, the variability of data inputs and building characteristics between researches. A brief comparison between the models can be seen in Table 2.1.

Table 2.1 - Brief comparison of machine learning models used in building energy consumption [58]

Models	ANN	SVM	Ensemble
Advantages	- Solve complex nonlinear problems - In general, better performance prediction than SVM	- Good balance between prediction accuracy and calculation speed - Few parameters need to be determined	- Best prediction accuracy and stability
Disadvantages	- Many parameters need to be determined	- Kernel function is crucial and difficult to be determined	- Difficult to implement
Computational Speed	Medium speed	Medium Speed	Low speed
Accuracy	Good	Average	Best

Additionally, it is worth to mention, the increase use of other techniques in the studies reviewed, like statistical and evolutionary algorithms, to optimize performance (computational time and accuracy). The main use of evolutionary algorithms, such as PSO, GA and DE, are incident in the learning process and model parameters selection, examples of it may be found in [47] [49] [54], which revealed better results, in comparison with models without it. Apart from the evolutionary algorithms a Bayesian optimization was also used to elect the best parameters in [48].

Furthermore, statistical analysis, such as mutual information and correlation [45], was often used in the input data selection stage that appears to be one of the major tasks, since the performance of the models is highly dependable of the right selection of the input variables. The input variables chosen in each of the reviewed papers may be seen in Table 2.2.

Based on Table 2.2, the following conclusions can be made:

- Since nearly all the studies were predicting one hour or a day a-head building energy consumption, most of them use different ranges of previous hours or days of its own consumption. Which normally gives better accuracy prediction considering that works as a 'guideline' for the corresponding model;
- Among all the weather conditions, temperature and solar radiation are of great importance to this type of forecasting, followed by relative humidity. The other variables do not manifest great use in the revised studies, except in [47] and [49] where it was used a particle component analysis (PCA) with two components, to reduce the dimensionality of all available weather conditions;
- As regards of indoor conditions, the most significant input used is building occupancy that yields best accuracy values among all the studies: in [47] and [49] for ANN; and in [55] for SVM. Most of the times, occupancy is not available for service buildings, although it is of great use once energy consumption is highly related to building daily inhabitants and their habits;
- In terms of calendar data, inputs such as hour and day type are used in most of the studies, followed by weekday that is often used. The reason why hour and weekday are selected is due to the consumption seasonality that may be found within daily and weekly time intervals. Additionally, day type is chosen because of the disparities between workday and non-workday daily consumption patterns.

Table 2.2 - Distribution of the selected input variables among the studies reviewed

Models	ANN	SVM	Ensemble
Energy Consumption			
Building previous energy consumption	[37], [43], [45], [46], [47], [49]	[53], [55]	[59]
Weather conditions			
Dry-bulb outdoor temperature	[37], [39], [43], [45], [46], [48], [47], [49]	[52], [53], [54], [55]	[60], [61], [59]
Relative humidity	[46], [48], [47], [49]	[52], [54], [55]	[60], [61]
Solar radiation	[39], [45], [47], [49]	[52] , [54], [55]	[61]
Dew point Temperature		[53]	[60]
Wind speed	[47], [49]		[60], [61]
Barometric pressure			[61]
Precipitation			[61]
Indoor Conditions			
Building occupancy	[39], [47], [49]	[55]	[60], [61]
Temperature	[46]		
Other building components (e.g. HVAC)	[46], [48]		
Calendar Data			
Hour	[37], [39], [45], [47], [49]	[55]	[61]
Weekday	[39], [48]	[55]	[61]
Month		[55]	
Day type (working/non-working)	[37], [43], [45], [48]	[53], [55]	[61]

From this point forward, the input variables that feed the models, are going to be referred as features, which is a term normally used in the machine learning field.

2.3 Complementary Models

To complement this study literature, it is mandatory to talk about other models that may have an important role in the performance of regression machine learning algorithms, mention in section 2.2.2

The complementary models can be grouped according to their functionality:

Data imputation - is when a model is used to impute missing values that, in the absence of it are lost. This type of procedure is applied in circumstances where the available data has a large amount of unknown values which may affect drastically the regression machine learning model performance.

This process can be done by several statistical or machine learning methods. The challenge remains in choosing the one that best fits the type of missing data. These methods can be applied in two distinct approaches, univariate or multivariate. The univariate approach is when a chosen method uses only the feature with the missing data to proceed to the imputation. Alternatively, in a multivariate approach, the use of other features is allowed, which in a data set where the features are related with each other, results in enhanced imputation predictions. The data imputation models presented below are within the multivariate approach.

- **Multivariate imputation by chained equations** (MICE) - as a classical approach for data imputation, basically works by filling the missing data multiple times. This is known as multiple imputation technique which although more computational expensive, performs better than single imputation as it measures the uncertainty of the missing values more accurately. The chained equations approach can be very flexible handling with different types of missing features in the same data set. This model can be broken down into four main steps.

  In the first step, a simple mean imputation is temporarily performed in every unknown value of the data set. Afterwards, in the second step, one of the features mean imputation values are set back to missing. Thirdly, the values from the feature chosen in second step work as the dependent variable in a regression model and all other features as independent. The regression model can be linear or logistic depending respectively on the chosen feature data type, numerical or categorical. In the fourth step, the values obtained from that regression are imputed, filling any unknown value of the chosen feature. In the following step, the cycle from the second step through the fourth is repeated for each feature, completing one iteration. The number of iterations may be pre-established by the user. At the end, each feature missing values have been replaced by the predictions of the performed regression models [64].

- **Miss Forest** (MF) - imputes missing values using the multiple imputation technique referred in MICE, but now with the RF model mentioned in 2.2.2.

  By default, the model starts by imputing missing values in the feature with the smallest number of unknown values, denoted as the candidate feature. Afterwards, any missing value of the remaining non-candidate features are filled by the mean or the mode, when it is a quantitative or a qualitative feature, respectively. With all the data set completed it proceeds to the next step that is characterized by the application of the RF. In this step the non-candidate features work as input data to the model and the candidate feature as output. The predicted values from the model are then used to replace the unknown values in the candidate feature. Following this, the imputer moves to the next candidate that is chosen with the same

criteria as before (feature with lowest number of missing values). The process repeats itself for each feature with missing values, until absence of missing data [65].

The latter model, shown outperforming results in diverse types of data sets when compared with the classical approach, MICE [65].

Data aggregation - the model presented here is in the scope of unsupervised clustering machine learning models referred in 2.2.2, and as said before, has the aim of grouping data points with similar characteristics.

- **K-means -** is used to group similar data points and discover underlying patterns. To achieve this objective, k-means looks for a fixed number of clusters (k) in a dataset, predefined. The k number, refers to the number of centroids, that represent the centre of each cluster. This algorithm identifies centre k number of centroids, and then allocates each data point to the nearest cluster through reducing the in-cluster sum of squares, keeping the cluster as small as possible. The learning process is performed iteratively to optimise the location of the chosen centroids, Figure 2.7. The algorithm will terminate if the iterations are maximized or if the centroids stop moving.

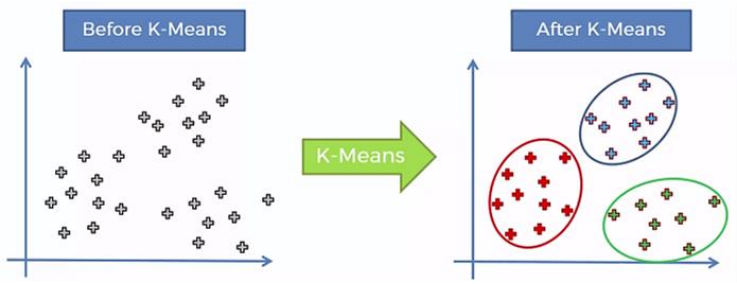


Figure 2.7 - Example of k-means clustering - with k = 3 [72]

To enhance this algorithm performance, it may be used a silhouette score and t-SNE algorithm mention in 2.2.2.

The silhouette score is a measure that quantifies how close each value in a cluster is to the values in its neighboring clusters. It is an elegant way to find out the optimum number of k during k-means clustering. This score lies in a range of [-1, 1]. A score of -1 indicates that the value is close to its neighboring cluster than to the cluster it is assigned to. Contrarily, if the score is +1 the value is far away from its neighboring cluster and extremely close to the cluster it belongs. In addition, if the score lies in-between the range, 0, it means that the value is at the boundary of the distance between the two clusters. Therefore, the higher the score the better is the cluster configuration. In order to get the best configuration possible, the algorithm is executed multiple times with different values of k. The one that provides the closest silhouette score to +1, is selected.

Additionally, the t-SNE algorithm may be used as a tool to visualize the distribution of the prior selected k from the silhouette score. This unsupervised dimensionality reduction algorithm has the ability of mapping high-dimensional space into a lower-dimensional order, this is done by calculating conditional probabilities between each space. The result of the algorithm supports the decision of the k clusters by creating a single map that reveals the structure of the analysed data.

2.4 Error Metrics

A crucial part of every data analysis project is the use of the proper error metrics to evaluate each model used. Several performances measure may be used in the forecasting of energy consumption, although the ones used in this work were selected based on the most frequently applied in the reviewed literature [26]. Therefore, the three error metrics used were: the coefficient of variation of the root mean square error (CV(RMSE)) (2.1), the mean absolute percentage error (MAPE) (2.2), and the mean absolute error (MAE) (2.3).

$$CV(RMSE)\ (\%) = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_{predict,i} - y_{true,i})^2}{n}}}{\bar{y}_{true}} \times 100 \tag{2.1}$$

$$MAPE\ (\%) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_{predict,i} - y_{true,i}}{y_{true,i}}\right| \times 100 \tag{2.2}$$

$$MAE\ = \frac{1}{n}\sum_{i=1}^{n}\left|y_{predict,i} - y_{true,i}\right| \tag{2.3}$$

Where $y_{predict,i}$ is the predicted energy consumption at time point i, $y_{true,i}$ is the actual energy consumption at time point i, y_{true} is the average energy consumption, and n is the total number of data points in the dataset.

From all the error metrics, CV(RMSE) was the most used evaluation measure in most of the studies performed. This may be explained by the fact that it is one of the performance evaluation measures recommended by the ASHRAE for this type of problems evaluation and also due to its comparison capabilities between different buildings, achieved by the normalization of the energy consumption prediction error with the average energy consumption.

Chapter 3

Study Case

3.1 Building Introduction

In the scope of this study, four buildings from Alameda campus of Instituto Superior Técnico (IST), Lisbon, were analyzed, namely, Civil, Central, North tower, and South tower buildings. In Figure 3.1 is illustrates a map of Alameda campus where is enumerated the location of each building.

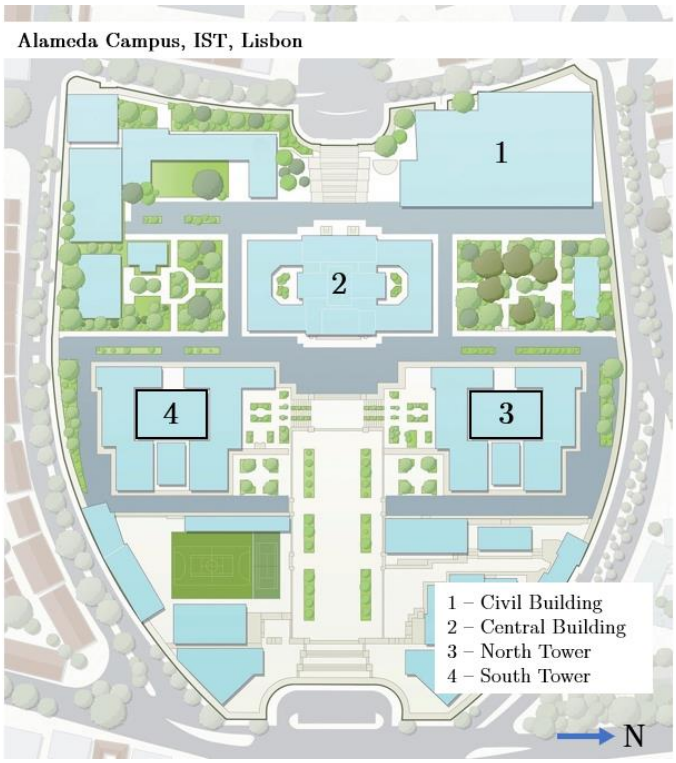


Figure 3.1 - Alameda campus, IST, Lisbon

Each building, as an independent system, has its own characteristics and main features, due to that each building was analyzed individually. In a way to simplify these analyses, it is mandatory to identify the patterns that dynamically affect building's energy consumption. These patterns may rely on key factors that at first sight are critical to energy use, such as occupancy, weather conditions and structural characteristics.

The stochastic behaviour of the occupants through their presence and activities in the building, influences energy consumption, not only passively by their metabolic heat produced, but also actively by their energy use (e.g. use of hot water, electrical appliances, lighting, and building openings) which results in an increase of internal heat gains and energy consumption, respectively.

Additionally, weather conditions parameters, such as outdoor temperature, relative humidity, and solar radiation are also important factors that may impact building's indoor thermal comfort and subsequently its inhabitants, mostly due to their annual variation. Thermal comfort can be reached through the proper use of the HVAC system, which is one of the major contributors in terms of energy use [8].

Furthermore, buildings' structure characteristics is another factor that may affect the indoor environment and consequently, energy consumption. Diverse type of structures and construction materials impact negatively or positively buildings' overall thermal properties. In a well-conceived building, its structure is used as an energy storage medium that supports building thermal management, by adapting to the various weather conditions scenarios.

To distinguish the different patterns caused by the earlier mention factors, each building energy consumption was evaluated in monthly, weekly and daily temporal granularities, with the data acquired from the previous year of the forecast, 2017. To conduct this visualization analysis, two different tools were used: boxplots and the k-means algorithm presented in 2.3.

3.2 Buildings Overview

3.2.1 Main Characteristics

From the four buildings presented earlier, it is possible to distinguish certain similarities since they are all institutional facilities, and also several particularities once the range of requirements of energy use in university campus is quite vast.

It was shown from the last energy audits [73]**,** that the main forms of energy use to fulfil their needs are natural gas and electricity. Electricity covers all the demands as regards to lighting, computers, plug-in devices, catering, common systems and laboratories facilities. In addition, natural gas is mainly used for spaces under concession and certain laboratories facilities. Besides that, HVAC system in each building, operates with different forms of energy in order to supply heating and cooling loads, in cooling season (May, June, July, August, September, and October) and heating season (rest of the months), respectively. **Central** and **Civil** building use electricity in both heating and cooling seasons, while **South** and **North Tower** use natural gas and electricity for cooling and heating seasons, respectively. It is also important to referred that a significant part of each building is designed for educational purposes including: classrooms, amphitheaters, laboratories, and other facilities; the remaining area is used for spaces under concession, except for **Civil** building that also has a data center.

Besides their similarities, they have different structures. **Civil** building, with a total floor area of 25.152 m^2, consists of 7 floors, of which 3 are underground (03, 02, and 01), and 4 aboveground (0, 1, 2, and 3). **Central** building, with a total floor area of 10.991 m^2, has 4 floors, concretely, 01 (underground), 0, 1, and 2. **North tower** and **South tower** have symmetrical structures, with a total area of 10.100 m^2 each, consisting on 16 floors, where 5 are underground (from 03 to 1 level), and the rest are raised

(from 2 till 12 level). Among the 16 levels, 4 are used only for technical purposes, namely, 03, 2, 3, and 12 floors. They also have different operating hours. **Civil** building is open during the week from 7 am till 9 pm and on Saturday from 7 am to 5 pm, despite that, in floor 0, it has a study area which is open 24 hours a day. **Central** building opens from 7 am to 9 pm on weekdays and from 7 am to 5 pm on Saturdays. **North tower** and **South tower** also have the same operation schedule, ranging from 7 am to 8 pm during weekdays and being publicly closed on Saturdays. In Table 3.1 it is shown a summary of each building main characteristics.

Table 3.1 - Summary of each building characteristics

Building		Civil	Central	North Tower	South Tower
HVAC energy use	Cooling season	Electricity	Electricity	Natural Gas	Natural Gas
	Heating season	Electricity	Electricity	Electricity	Electricity
Total floor area (m^2)		25.152	10.991	10.100	10.100
Total floors		7	4	16	16
Operation Schedule	Weekdays	7am-9pm	7am-10:30pm	7am-8pm	7am-8pm
	Saturday	7am-5pm	7am-3:30pm	-	-
24 hours spaces	Data center	No	Yes	No	No
	Study area	Yes	No	No	No

3.2.2 Energy Consumption Analysis

Monthly Analysis

For this analysis, the yearly energy consumption of the weekdays was divided in daytime (from 8 am till 8 pm) and nighttime (from 9 pm till 7 am), Figure 3.2 and Figure 3.3, respectively. In addition, it is also presented a boxplot for each building, without any filter applied, Figure 3.4.

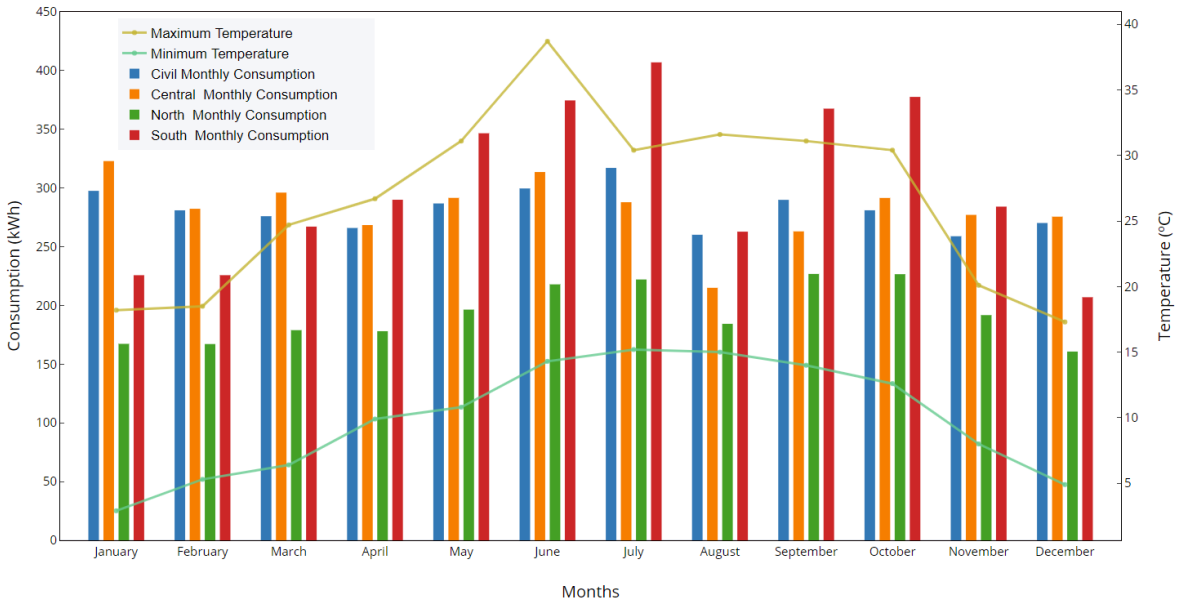


Figure 3.2 - Monthly mean weekdays energy consumption of each building, during the day (from 8 am till 8 pm)

In general, during the daytime, Figure 3.2, it is possible to see a higher mean consumption tendency, during the cooling season when compared with heating season. This tendency is highly related to

HVAC system energy use to maintain the indoor thermal comfort, in the presence of more adverse conditions, which occurs in summer. Although there is an exception of this tendency, in August, that shows the lowest mean energy consumption when compared with the rest of the months during the year, caused by the two weeks summer break of all campus facilities, evident in Figure 3.4. Moreover, it is noticeable that **North** and **South tower** are more exposed to the outside conditions than **Central** and **Civil** buildings, achieving higher mean consumptions in cooling season, this is mostly due to the thermal properties of the glazed facades that surround both **towers**, which are more vulnerable to solar energy. In both **towers** it is also perceptible that the consumption fluctuation follows the maximum and minimum temperature variations along the year. This consumption curve is characteristic of buildings using natural gas and electricity for cooling and heating seasons, respectively. On contrary, **Civil** and **Central** building, present a more stable consumption curve which is substantiated by the main use of electricity. These patterns are more clear represented in Figure 3.4.

Additionally, besides the existent structure symmetry between **North** and **South tower**, their consumption is distinct, being **South tower** the higher energy expenditure, Figure 3.2. One of the reasons that may be related to this expenditure is the type of existing research facilities that are mainly used for chemistry purposes with equipment that requires larger amounts of energy.

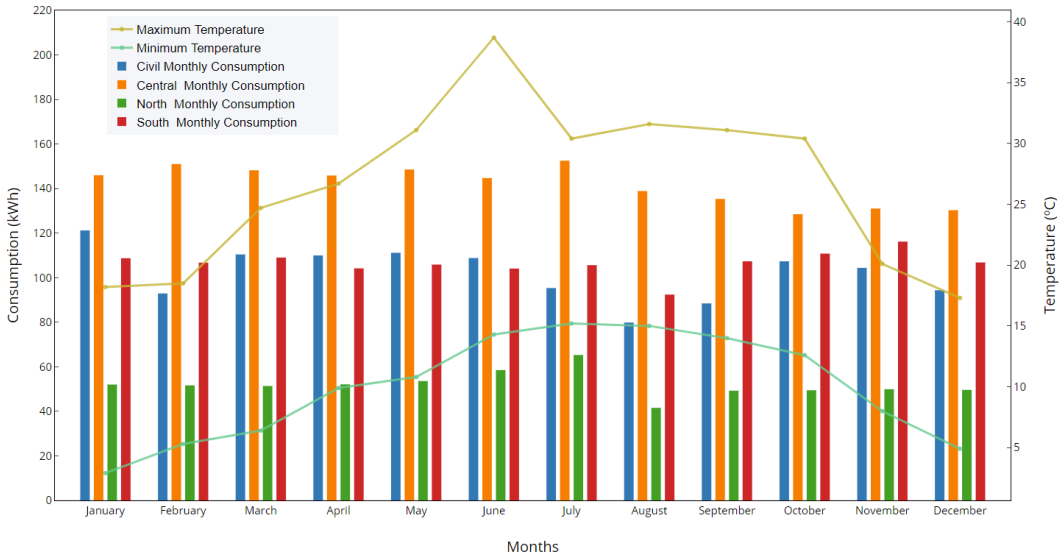


Figure 3.3 - Monthly mean weekdays energy consumption of each building, during the night (from 9 pm till 7 am)

During nighttime, Figure 3.3, two main behaviours are visible. In **Civil** building, the existence of a study area open 24 hours a day, creates a pattern with higher mean consumption in the months corresponding to the two semesters and evaluation periods, e.g. in 1º semester evaluation period, in January, that reveals a high mean energy expenditure. On the other hand, it is also clear that **Central** building has the largest mean consumption among the other buildings, which may be justified by the 24 hours operating data center.

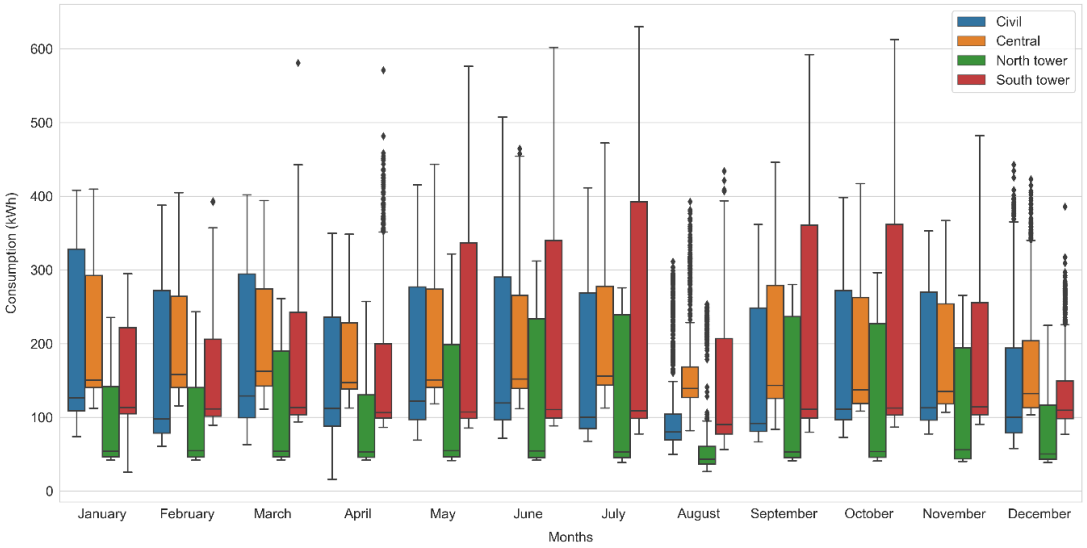


Figure 3.4 - Monthly boxplot energy consumption of each building

Weekly Analysis

In each building there is a 7 days cycle pattern that repeats almost through the whole year, as it is represented by the weekly boxplots in Figure 3.5. As expected, most of the buildings' energy use occurs during weekdays since it is when the majority of buildings' activities take place. It is also noticeable a slight decrease in energy expenditure in the last weekday, Friday, probably due to the arriving of weekend and people's tendency to leave earlier. During the weekend, there is an abrupt fall in energy consumption, although Saturday energy use is slightly higher than Sunday, due to weekend opening hours, scheduled in Table 3.1. Nevertheless, this pattern can be altered by local or national holidays that consequently may reduce buildings' occupancy rate resulting in less energy expenditure.

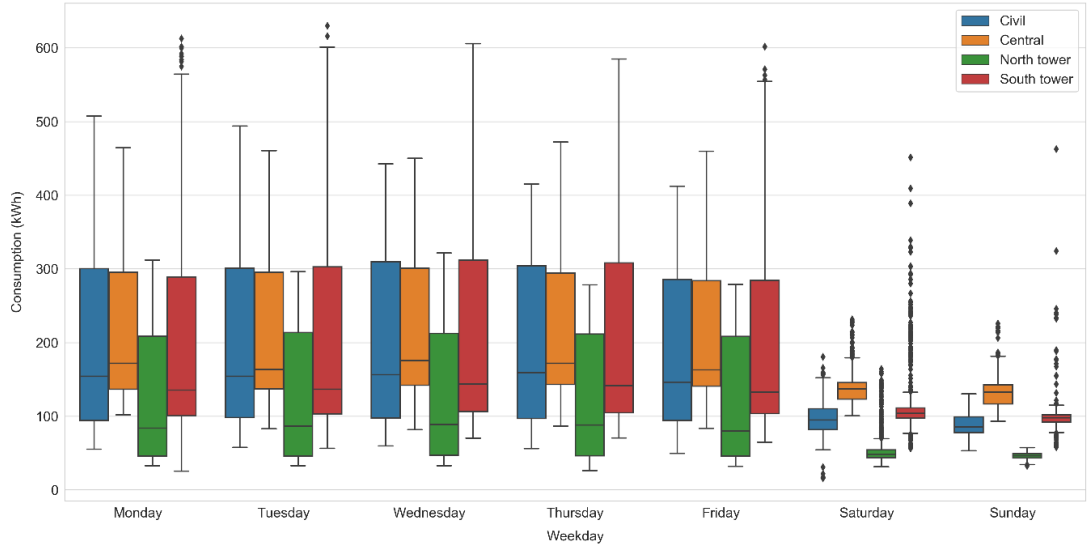


Figure 3.5 - Weekly boxplot energy consumption of each building

Hourly Analysis

In hourly analysis, each building was evaluated individually to simplify the visualization and gathered each different daily consumption curve characteristics.

To recognize the different daily consumption patterns, it was used the clustering algorithm, mention in 2.3. In this context, the algorithm groups the most similar days, based on the chosen number of clusters (k). To identify the right k, a range between 2 to 15 clusters was tested via the silhouette score, Figure 3.6, and the one selected was posteriorly represented by the t-SNE algorithm in a lower-dimensional space attached to the correspondent building daily consumption patterns, Figure 3.7.

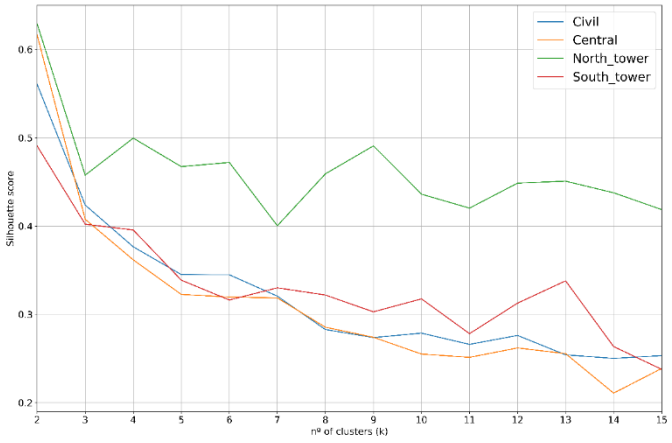


Figure 3.6 - Silhouette score of different number of clusters by building

From the silhouette scores, Figure 3.6, most of scores were indicating the use of two clusters per building, although in the scope of this analysis the evaluation started at k=3, in attempt to distinguish a typical non-working day and a typical working day during cooling and heating season. With that in mind, Civil and Central building maintain the k=3, once for both it was the next highest score. As regard to both towers, the selection of k was controversial. For North tower, despite that the highest score may be found with 4 clusters, it was chosen 3, since the use of more clusters identify mostly the hour shifting patterns, giving no additional information for this analysis. On contrary, for South tower, it was adopted k=4, since all the clusters represent distinct patterns, regardless the slightest lower score.

After this selection, the hourly mean consumption of each cluster was highlighted and a table with the percentage of days included in each cluster are distinguished by type and season, leaving all prepared to evaluate each building daily patterns.

As regards to **Civil** building, Figure 3.7, in general, the daily energy patterns are not influenced by heating and cooling seasons as referred earlier in monthly analysis, since every cluster has around 50% of the days in each season, as shown in Table 3.2.

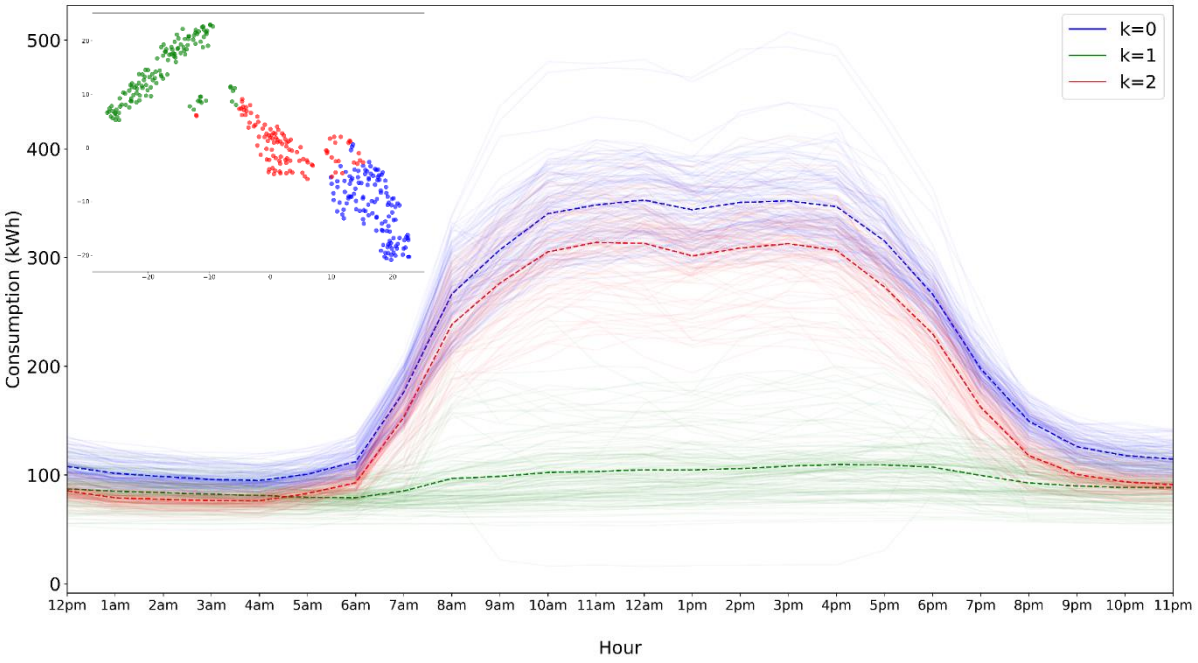


Figure 3.7 - Civil building daily consumption patterns defined by k-means algorithm (k=3) and t-SNE distribution

Table 3.2 - Civil building day type percentage of each daily consumption patterns defined by k-means algorithm (k=3)

Cluster	workday (%)	holiday (%)	weekend and summer break (%)	heating season (%)	cooling season (%)
k=0	100	0	0	55.8	44.2
k=1	3.1	11	85.9	47.7	52.3
k=2	98.1	1.9	0	43.9	56.1

In addition and based on Figure 3.7 and Table 3.2, the following is possible to be concluded:

- k = 0 and k = 2 - consisting mainly of working days represent two different typical workday consumption patterns. In terms of the highlighted average energy consumption patterns, in general, they start increasing at 6:30 am achieving values around 325 kWh before 12 am, afterwards they have a slightly consumption decrease corresponding to lunch break. During the afternoon the patterns are almost a morning mirror, decreasing to building base energy around 9 pm. Between both clusters, k = 0 represents only working days, achieving a higher mean energy consumption pattern than k = 2. The presence of holidays in the latter mention cluster, may be explained by the 24 hours studying area used during exams periods.

- k = 1 - includes holidays, summer break days and weekends, being the latter predominant. The mean consumption of those days is nearly stationary with a slightly higher expenditure during part of the daytime mostly due to Saturdays opening hours and the 24 hours studying area.

Similar to the previous building, **Central** building has also no significant influence of heating and cooling seasons in daily patterns, Table 3.3. In addition, the 24 hours data center results in a building base energy expenditure around 130 kWh, Figure 3.8, higher than the rest of the buildings.

Furthermore, and based on Figure 3.8 and Table 3.3, the following is possible to be concluded:

- k = 0 and k = 2 - like the previous building, are formed mostly by working days, with two typical mean consumption patterns that characterizes two groups of mainly working days. These patterns, start ascending at 6 am achieving a first peak at 11 am, then the consumption decrease till around 12 am which is when the lunchtime takes place having a roughly constant period during 1 hour, more obvious than the previous building. Afterwards, the average consumptions start rising again achieving the second peak values at 3 pm, from that point forward both clusters mean energy consumption decline till reach the building base energy consumption at 8 pm. Between both clusters, k = 2 representing not only working days but also summer break days, has a mean daily consume lower than k = 0, although the inclusion of those days in k = 2 most be related by atypical opening hours during summer break for maintenance.

- k = 1 - is mainly formed by holidays, summer break days and weekends, being the latter dominant. The mean consumption curve characterizing those days is constant, similar to **Civil** building, but with a higher mean energy consumption, probably due to the 24 hours data center.

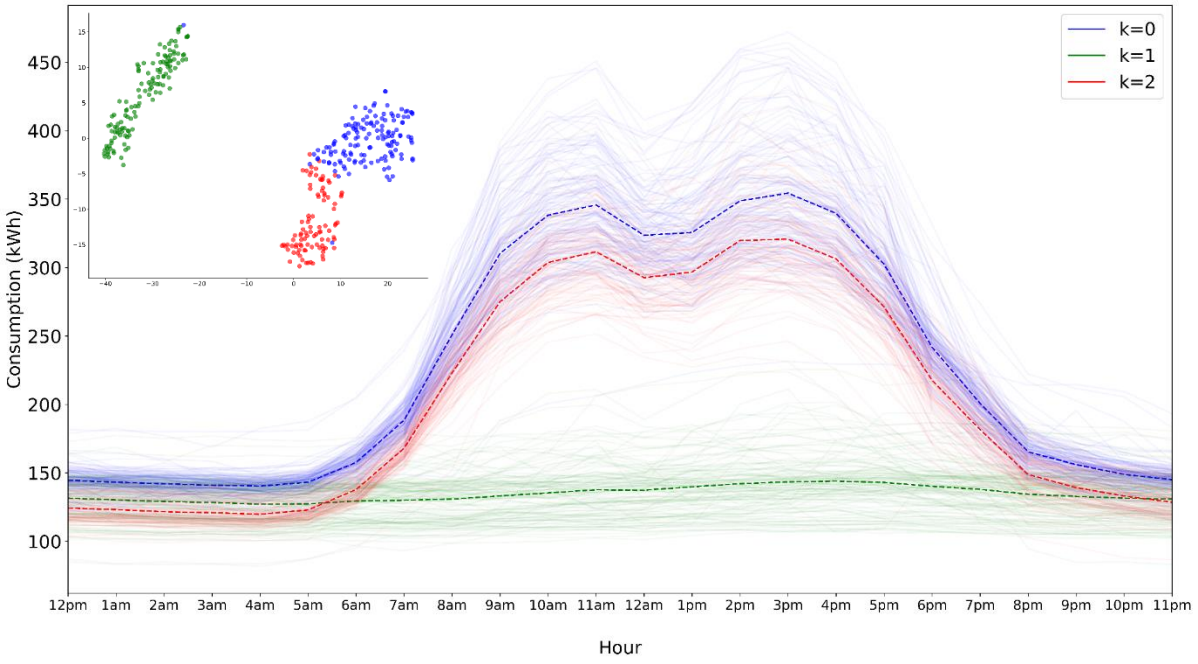


Figure 3.8 - Central building daily consumption patterns defined by k-means algorithm (k=3) and t-SNE distribution

Table 3.3 - Central building day type percentage of each daily consumption patterns defined by k-means algorithm (k=3)

Cluster	workday (%)	holiday (%)	weekend and summer break (%)	heating season (%)	cooling season (%)
k=0	99.3	0	0.7	48.2	51.8
k=1	2.3	12.5	85.1	46.9	53.1
k=2	100	0	0	54.7	45.3

As regards to **North tower** it is possible to mention its influence by heating and cooling seasons as referred earlier in the monthly analysis, once the working day clusters, k = 0 and k = 2, have greater percentage values of cooling season and heating season days, respectively, Table 3.4. It is also noticeable, the lack of the typical low consumption range that occurs in the previous buildings' lunchtime.

Based on Figure 3.9 and Table 3.4, the following is possible to conclude:

- k = 1 - consists of working days mainly during cooling season. The mean consumption begins at 6 am and rapidly achieves a peak at 8 am, due to the start operation of an HVAC's system component, named as chiller. Afterwards, it has a light positive slope during the day reaching values above 250 kWh at 3 pm, from that hour forward the energy expenditure decreases, till 9 pm, where it catches up building base energy consumption. In this cluster of days, it is particularly visible the shifting hour due to the abrupt energy increase and decrease in the beginning and ending of the day, respectively.

- k = 2 - is mainly formed by holidays, summer break days and weekends, being the latter more predominant. The average consumption is again almost constant, with a slight growth during the day, most certainly due to Saturdays.

- k = 0 - consists mainly of working days during heating season. From the average energy consumption pattern, it is perceptible the overall lower consumption in comparison to k = 1, this is mostly due to the less adverse weather conditions felt in winter than in summer. This average pattern starts at 6 am and rises to a peak at 10 am, afterwards the consumption fluctuations do not vary much till 4 pm, where it decays achieving the buildings base energy at 9 pm.

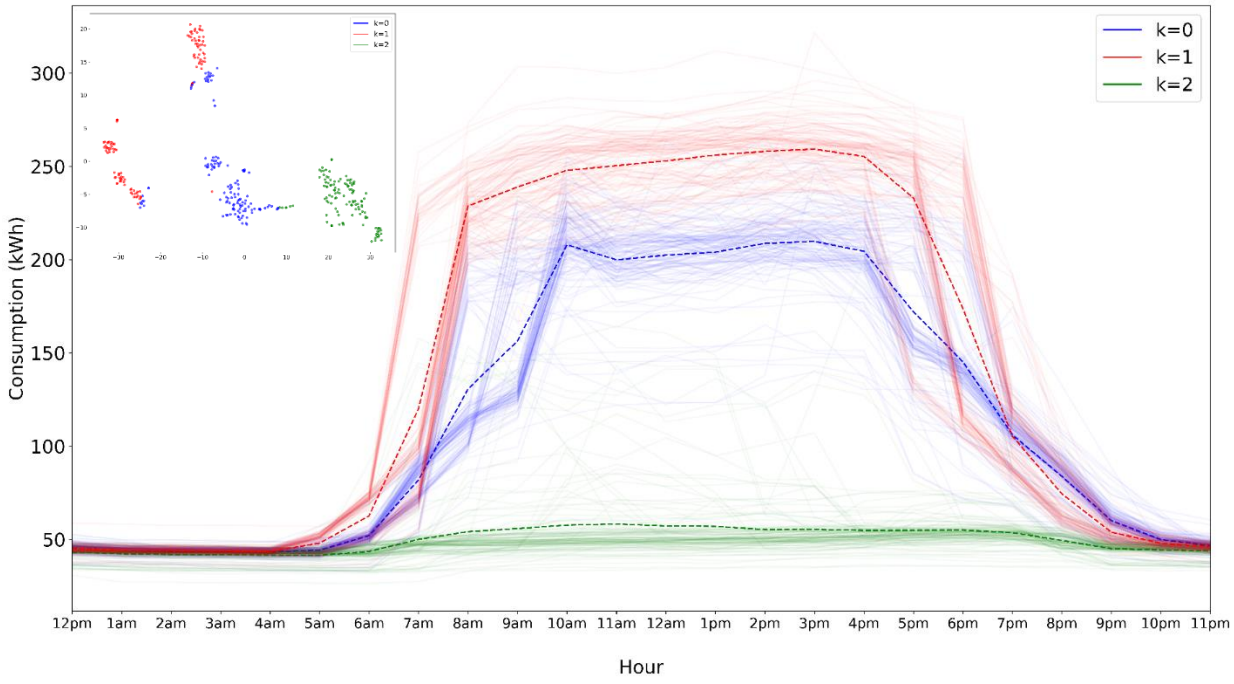


Figure 3.9 - North tower building daily consumption patterns defined by k-means algorithm (k=3) and t-SNE distribution

Table 3.4 - North tower building day type percentage of each daily consumption patterns defined by k-means algorithm (k=3)

Cluster	workday (%)	holiday (%)	weekend and summer break (%)	heating season (%)	cooling season (%)
k=0	93.8	5.4	0.8	88.5	11.5
k=1	100	0	0	8.8	91.2
k=2	1.7	7.5	90.8	45.9	54.1

Similar to the previous building, **South tower** is also affected by heating and cooling seasons in daily patterns, Table 3.5. According to the monthly and weekly analysis, between both towers this building is the one who expends more energy and that can be clearly visible when comparing the average consumption patterns between Figure 3.9 and Figure 3.10.

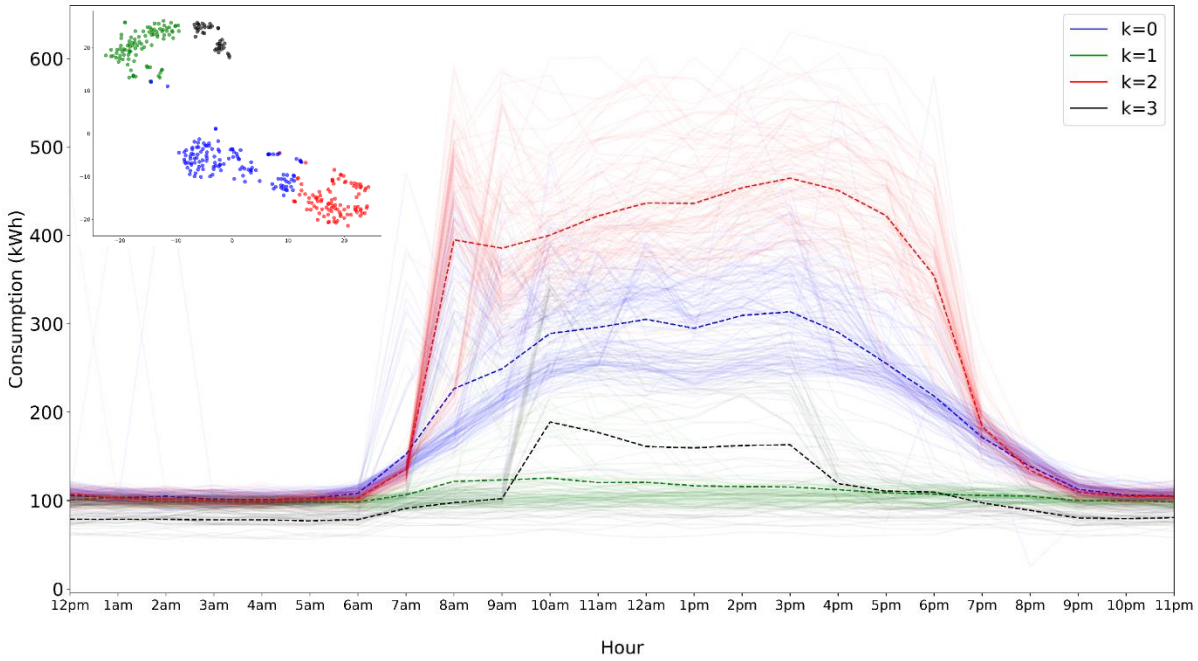


Figure 3.10 - South tower building daily consumption patterns defined by k-means algorithm (k=4) and t-SNE distribution

Table 3.5 - South tower building day type percentage of each daily consumption patterns defined by k-means algorithm (k=4)

Cluster	workday (%)	holiday (%)	weekend and summer break (%)	heating season (%)	cooling season (%)
k=0	96.2	2.3	1.5	84.7	15.3
k=1	9	9.8	81.2	49.6	50.4
k=2	100	0	0	3	97
k=3	27	8.2	64.8	24.4	75.6

Based on Figure 3.10 and Table 3.5, the following is possible to conclude:

- k = 0 - consisting predominantly of working days groups most of cooling season days. The mean energy consumption pattern during these days has no peculiar peak, starting at 6 am with a soft slope leading to the maximum energy expenditure around 300 kWh at 3 pm, after that consumption gently decays to base building energy at 9 pm. In between, is noticeable a slight consumption decreases during lunchtime.
- k = 1 - clusters principally holidays, summer break days and weekends, being the latter the dominant day type. The average consumption curve, as expected, behaves as almost a non-working day excepting an higher consumption happening from 9 am until 4 pm, which may be caused by the presence of working days in the cluster, accurately 9%.
- k = 2 - overall, is the cluster that exhibits the higher mean consumption pattern, most certainly due to the fact of grouping only working days during the cooling season. The mean energy

expenditure pattern starts rising at 6 am, and between 7 am and 8 am, ascends sharply to a first consumption peak, which can be again explained with the start of chillers operation. Subsequently, apart from the minor expenditure reductions at 9 am and 1 pm, the consumption keeps increasing to the higher energy peak at 3 pm which may be considered the time of the day where usually outside conditions are most extreme (high temperatures), demanding more energy of the HVAC system to counteract the situation and establish thermal comfort. This last peak can be also detected in **North tower** but not that evident. From 3 pm forward the energy expenditure decays, with an abrupt fall during 6 to 7 pm and reaching building base energy consumption at 9pm.

- k = 3 - as a similar consumption to k = 1 group in terms of consumption, except between 9 am and 4 pm, where there is a slightest energy consumption increase.

Chapter 4

Methodology

After knowing the main characteristics of each building, the present chapter will demonstrate the methodology behind the development of this work to achieve its main goal: the energy consumption prediction of four buildings (Civil, Central, North tower, and South tower) in three different forecasting horizons (an hour, a day, and a week). A detailed diagram with each main step may be seen in Figure 4.1. To conduct each step, a programming language, known as python was used [66].

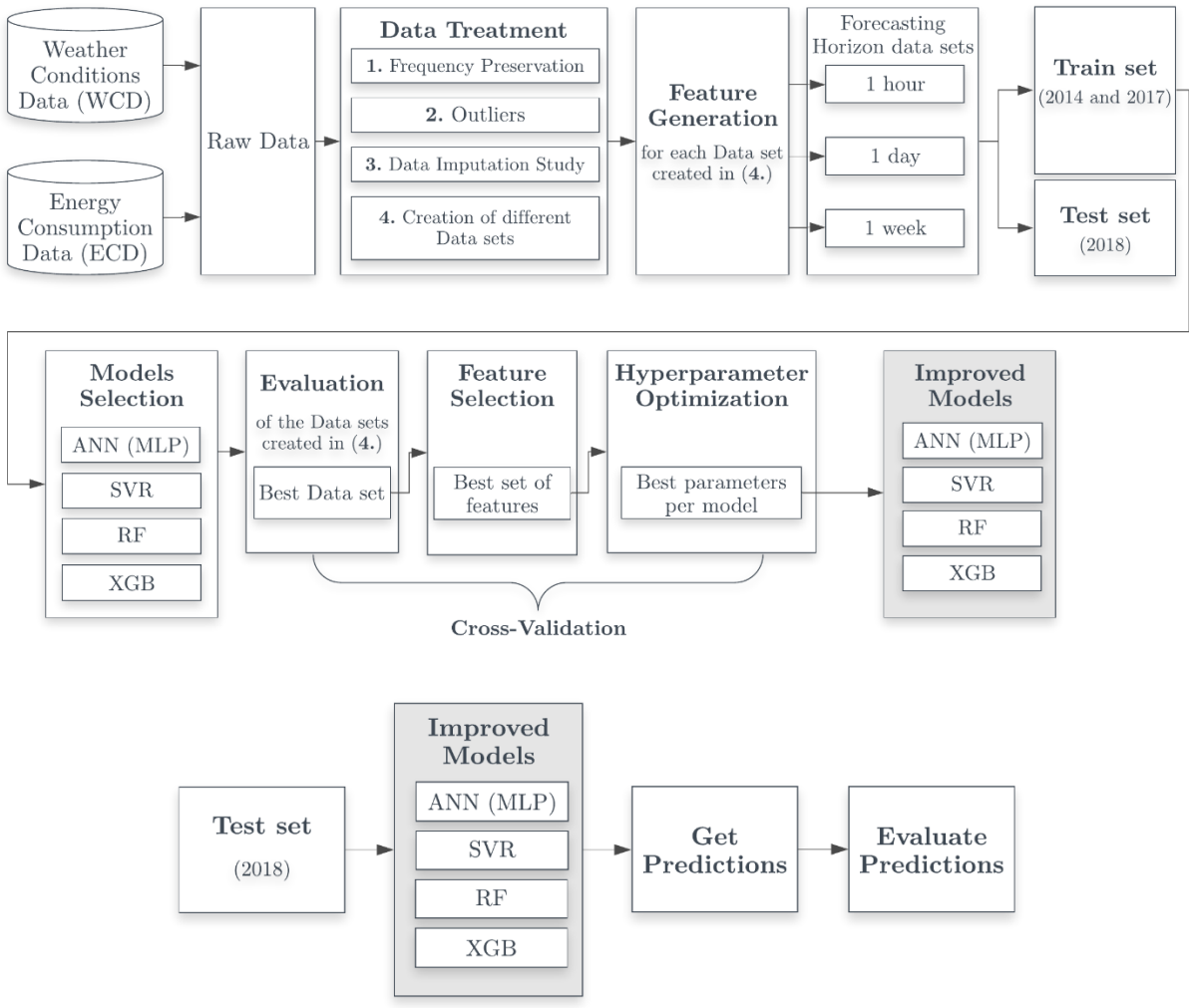


Figure 4.1 - Methodology step-by-step diagram

4.1 Data Treatment

The data set available in this work may be grouped into two different categories. The first category, named as the energy consumption data (ECD), contains values of each building energy consumption, collected at 1-hour intervals for 3 years (2014, 2017, and 2018). The second category, denoted as weather conditions data (WCD), includes the outdoor weather conditions, such as temperature, relative humidity, and solar radiation. This category was also hourly gathered during the same years by an existent weather station in Alameda campus.

Furthermore, the real-world data sets that are gathered from sensors and electric meters, often called as raw data, includes incomplete, imprecise and noisy values. To proceed to the model development, it is mandatory to acknowledge and handle the uncertainty of it. The process that addresses these issues is referred in this study as data treatment and may be branched into four sequential steps:

1. Frequency Preservation;
2. Outliers;
3. Data Imputation Study;
4. Creation of different data sets.

Frequency Preservation

In the first step, since data is time dependent, an hourly frequency preservation was done to keep the data set continuity. Therefore, every missing or repeated hour, was added or deleted, respectively, creating a missing value row if needed.

Outliers

Following that, in the second step, an outlier detection was performed in ECD. An outlier is an abnormal data value that considerably diverges from the rest of the data points in the same feature, their inclusion may affect negatively the predictive model accuracy. Although, in certain situations, an outlier may contain important information about a specific system behaviour. As a result of this ambiguity, two successive methods were applied. Firstly, a statistical metric, named as z-score was select to detect the outliers. This statistical metric quantifies every value of a given feature, by scoring is standard deviation from the mean. It is ruled by equation (4.1), where μ is the selected data mean, σ the standard deviation and x the data value that is scored.

$$z = (x - \mu)/\sigma \tag{4.1}$$

After the scoring, the values that presented an absolute standard deviation above or equal four ($|z| \geq 4$) were identify as outliers. Four was chosen as sigma, to guarantee that the selection of outliers was not too restrict. Afterwards, the second method starts by visualizing each of the days where the outliers were detected. This method works as a filter, to identify if any of the previous detected values correspond to an unexpected spike or dip during the day. If the latter statement occurs the outlier is replaced by a missing value in the data set. As a result of this analysis, just three values from South tower were set as missing values. In Figure 4.2, it is possible to see an example of an outlier that did not pass in both methods.

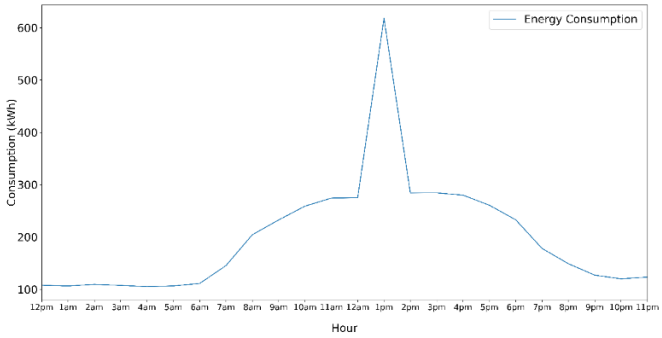


Figure 4.2 - South tower energy consumption - removed outlier, at 5pm, with z = 4

As a consequence of the first two steps, the data set presented a greater number of missing values than the original one. To proceed to the following step, it is mandatory to quantify those values and check their distribution per feature, with that end, Table 4.1 and Figure 4.3 are displayed. It is also worth to mention, that from this point forward the abbreviations shown in Table 4.1 will be used to refer to each feature.

Table 4.1 - Data set of the available features and their missing values

Available Features	Units	Abbreviation	Total values	Missing Values	
				(%)	Total
Buildings energy consumption data					
Civil	$[kWh]$	$civil$	26280	0.60	157
Central		$central$		0.02	4
North tower		$north\_tower$		0.32	83
South tower		$south\_tower$		0.36	95
Outside weather conditions data					
Temperature	$[ºC]$	wt_temp	26280	8.97	2357
Apparent temperature		wt_tmpap			
Relative humidity	-	wt_hr			
Mean wind speed	$[m/s]$	$wt\_mean\_windspd$			
Maximum wind gust		$wt\_max\_windgust$			
Mean atmospheric pressure	$[mbar]$	wt_mean_pres			
Mean solar radiation	$[W/m^2]$	$wt\_mean\_solarrad$			
Precipitation	$[mm/hr]$	wt_rain_day			

Furthermore, it is important to understand the reason why data goes missing. The major part of the discontinuity revealed in Figure 4.3 is due to sensors and electric meters malfunctions, although an exception occurred in 2018, during almost 2 months (1165 consecutive values) where it can be found a wide gap in the WCD. This exception is related with the waiting period between the weather station computer failure and the arrival of a new one.

Data Imputation Study

The third step focuses on the type of imputation applied to fill those missing values. As it is known, there is no good way of dealing with missing data, in this work data set the main percentage and resulting distribution of this lack of data was found in WCD, Table 4.1 and Figure 4.3. For this reason,

merely adopting a strategy of dropping those values will lead to a loss of the correspondent (same rows) ECD. Since the latter, contains the most valuable feature of each building, its own consumption, it is of extreme importance not losing it.

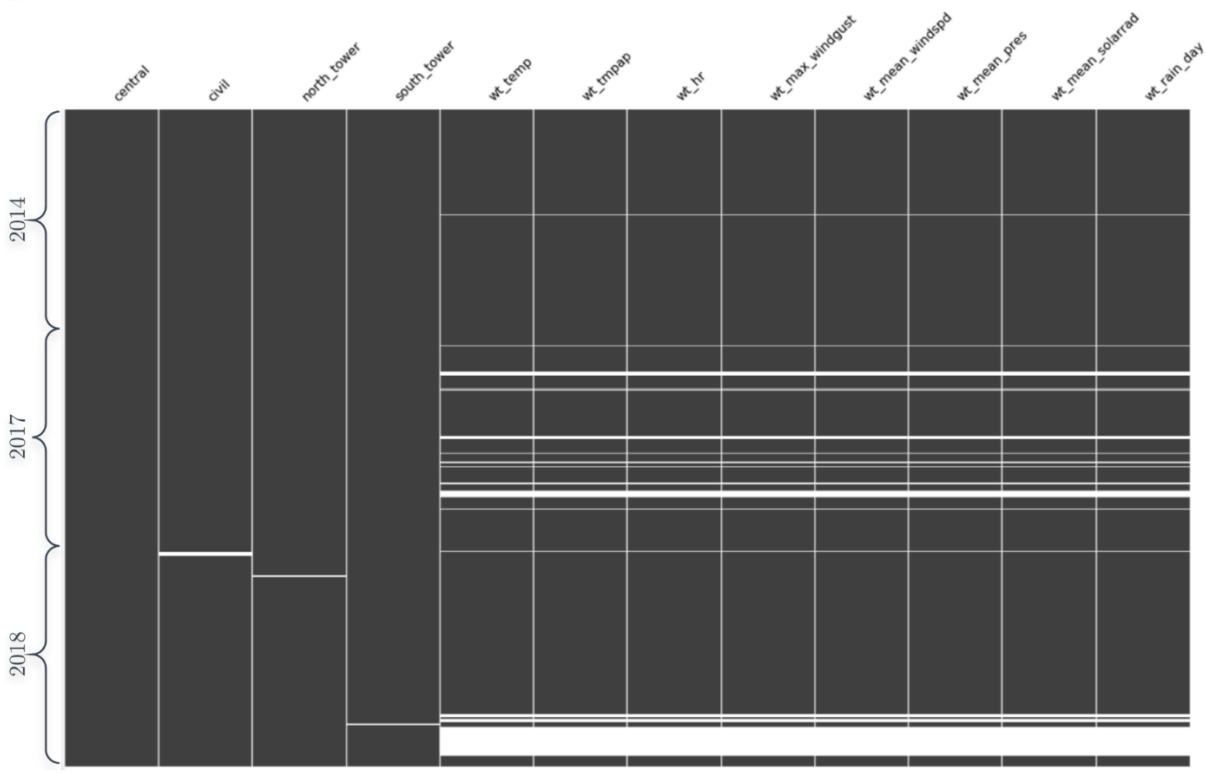


Figure 4.3 - Data set missing values distribution per feature

To overcome this situation, a study has been conducted to measure the accuracy of different imputation strategies. Two models have been used in this study, specifically MICE and MF, mention in 2.3.

To conduct this study, the process represented in Figure 4.4, was performed. This process was done separately for ECD and WCD, since both models are ruled by the multivariate approach which may lead to an exchange of information between each data set type and a consequent influence in the further performed forecasting. This process can be broken down into five main steps:

1. From the entire data set, a selection of the maximum consecutive missing values of each feature was performed, denoted as **target,** Table 4.2. This selection was motivated by the increased difficulty of imputing wide ranges of missing values.
2. The data from the year of 2014 was chosen, since it exhibits the lowest number of missing values, as can be seen in Figure 4.3. Any unknown value from the adopted year was dropped, and the rest of data was used as a baseline data set to test the different imputation models.
3. A random generation of artificial missing data based on the **target** value of each feature was performed in the baseline data set. Regardless the randomness of this part of the process, when the WCD was evaluated, the gaps are generated in parallel, in attempt to replicate the real situation, Figure 4.3.

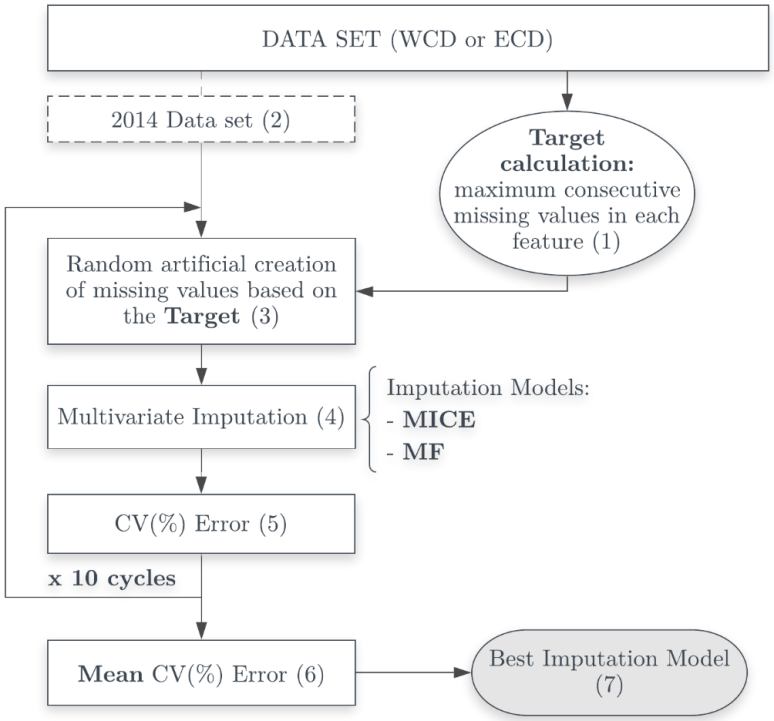


Figure 4.4 - Data imputation study step-by-step diagram

Table 4.2 - Target of each feature - maximum consecutive missing values in 2017 and 2018 data sets

Features	Target
civil	148
central	1
north_tower	76
south_tower	88
wt_temp	1165
wt_tmpap	
wt_hr	
wt_mean_windspd	
wt_max_windgust	
wt_mean_pres	
wt_mean_solarrad	
wt_rain_day	

4. Both models, MICE and MF, were separately applied, filling the missing values with the correspondent predictions.
5. The imputed values were then compared with the real ones, stored in step 3. The evaluation of the models per feature was conducted using the CV(RMSE) metric, mention in 2.4. Afterwards, the baseline data set predictions were set back to the real values.
6. Steps 3 to 4 were repeated for 10 cycles to ensure the diversity of the times of the year where the imputation was performed by each model. At the end of the cycles, the mean error was calculated per feature to evaluate the models applied.

7. A comparison of each model performance by feature was done. The RF model outperformed the other model in both ECD and WCD, showing superior accuracy in every feature, in terms of mean CV(RMSE).

After this study completion, although the MF model showed, in general, a great accuracy in each of the data set types when compared with MICE, the gap in WCD was still with undesirable imputation values in terms of the expected seasonality and trend.

To address this problem, a method based in the univariate approach, referred in this work as hour monthly mean (HMM), was developed. This method basically fills each feature independently with the known values from the other years, using the correspondent mean value of the same month and hour. For example, if a certain feature as a missing value at one day of August at 10 am, this method will use the mean of the other years known values of August at 10 am for the imputation.

A comparison of the two models and the new method imputation was done for each of WCD feature. An example of it, for wt_temp feature, can be seen in Figure 4.5, where it is possible to conclude that although none of the imputation strategies represents well the actual temperature, the new method gives the daily seasonality and the monthly tendency needed in comparison to the nearly constant values imputed by the MF and MICE.

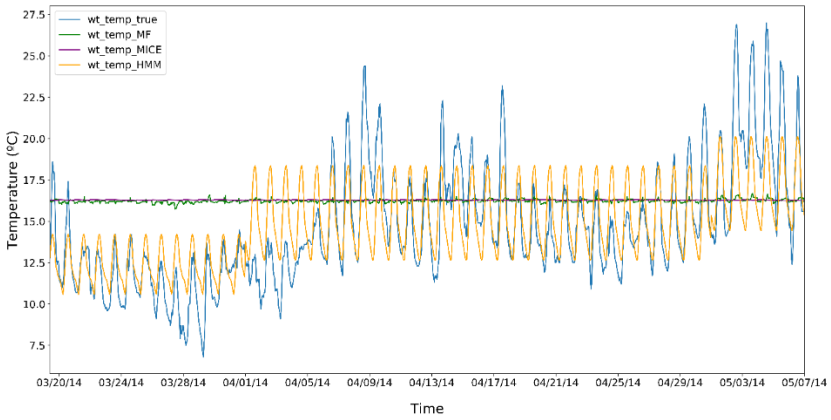


Figure 4.5 - Imputation comparison example of MF, MICE, and HMM imputations with the true value of wt_temp

Creation of different data sets

The fourth and last step, consists in the creation of three different data sets, that differ in the strategies adopted for the missing values shown in Figure *4.3* and Table *4.1*.

In this step, it was introduced another technique of univariate imputation, named as linear interpolation. This statistical technique essentially draws a straight line between two or more known values filling the gap created by the missing values. Due to its simplicity is of great importance to limit the range of unknown values to fill in order to avoid the replacement of a possible feature pattern with a straight line, with no relevant information. The limit applied was of three hours, meaning that if any of the features has three consecutive missing values a linear interpolation is applied between the value before and after the gap. The target of this technique was the sudden spikes or dips identified previously as outliers.

Therefore, the first strategy common to each of the data sets, named as dt_01, dt_02, and dt_03, was the three hours linear interpolation. After that, the three data sets differ from one to another, as it is described below:

- In dt_01 the remaining missing values were dropped, making it the least influenced by imputation;
- In dt_02 was applied a MF imputation just in the ECD. This was done to check the relevance of the imputation only in building energy consumption, since their error metrics shown the best mean results among the other features. Afterwards, the rest of the missing values were dropped.

In attempt to not losing any part of the data, the following data set was created:

- In dt_03 it was used the MF imputation model in ECD. After that, it was implemented the HMM method to the WCD, reasoned by the outcome shown in Figure 4.5.

In Table 4.3, it is possible to check the different techniques applied in each data set and the correspondent lost rows.

Table 4.3 - Summary of the imputation techniques applied per data set

Data sets	Imputation			Drop remaining missing values	Lost rows
	Linear Interpolation	ECD MF	WCD HMM		
dt_01	✔			✔	2590
dt_02	✔	✔		✔	2286
dt_03	✔	✔	✔		0

4.2 Feature Generation

The feature generation step is responsible to created new features that somehow offer additional information to the machine learning models towards the enhancement of their output predictions. In the scope of this work, the output is the energy consumption of each building, therefore, each new feature most be generated in attempt to describe its behavior and particular characteristics.

In this study, the new features, may be grouped according to what they are based on, therefore, there are three main categories, time, calendar, and energy consumption. The first includes every new feature that was time dependent, the second group of features was based on the national and academic calendar, and lastly, the third group generates features that were based on each building own consumption, Figure 4.6.

Time

This group of new features, was supported by the patterns encounter in each of the temporal partitions performed in the consumption analysis of 3.2.2. In attempt to offer to the forecasting models the possibility to distinguish those patterns three features were generated with integers values that differ depending on the temporal partition, e.g. for monthly partition, integers from 1 to 12 were set in the new feature. The three new features were named after each temporal partition, specifically t_month, $t\_dayofweek$, and t_hour.

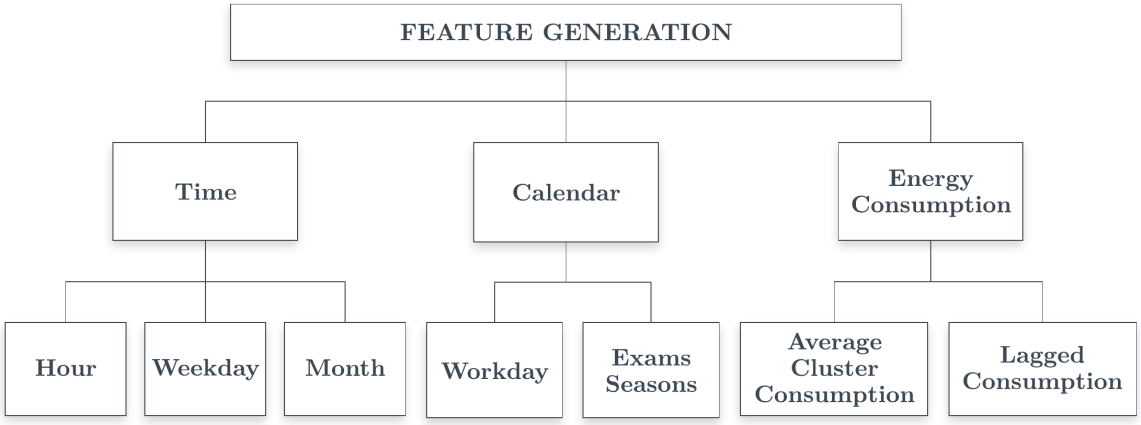


Figure 4.6 - New features categories

Calendar

In the absence of each building real occupancy data, this category was implemented. It basically attempts to replicate the real occupancy by day, with two different features. For that, each of the features use integer values, levels, that quantify the expected daily occupancy rate, being the lowest level the one that among the others has the smallest occupancy rate.

The first feature, named as $s\_workday$, specifies the type of day in three distinguished levels, defined as:

- **Level 0** - represents weekends and the two weeks yearly summer break of all campus facilities;
- **Level 1** - identifies the holidays that occur during the week. This level, although representing non-working days, in the case of an institutional buildings, their rate of occupancy is, usually, in-between the other two levels;
- **Level 2** - as the last level in $s\_workday$ feature, represents the normal working days, where the occupancy rate is expected to be the highest.

The second feature, denoted as s_epochs, attempts to give the model information about the different types of occupancy occurring when there is exams, classes, and break periods during the academic calendar. For that, also three levels were specified:

- **Level 0** - corresponds again, to the lowest rate of occupancy, the break period between semesters;
- **Level 1** - identifies the exams period, with no classes;
- **Level 2** - refers to the classes period of each of the academic semesters.

Energy consumption

This category was used to create a sort of “guidelines” using each building consumption, to enhance the performance of the forecasting models. These “guidelines” can be distinguished into two different groups, according to the technique used to generate them.

The first group used the mean daily pattern of every identified cluster in consumption daily analysis per building (3.2.2), to create a feature that repetitively replicates that pattern along all data set, as it is illustrate in Figure 4.7.

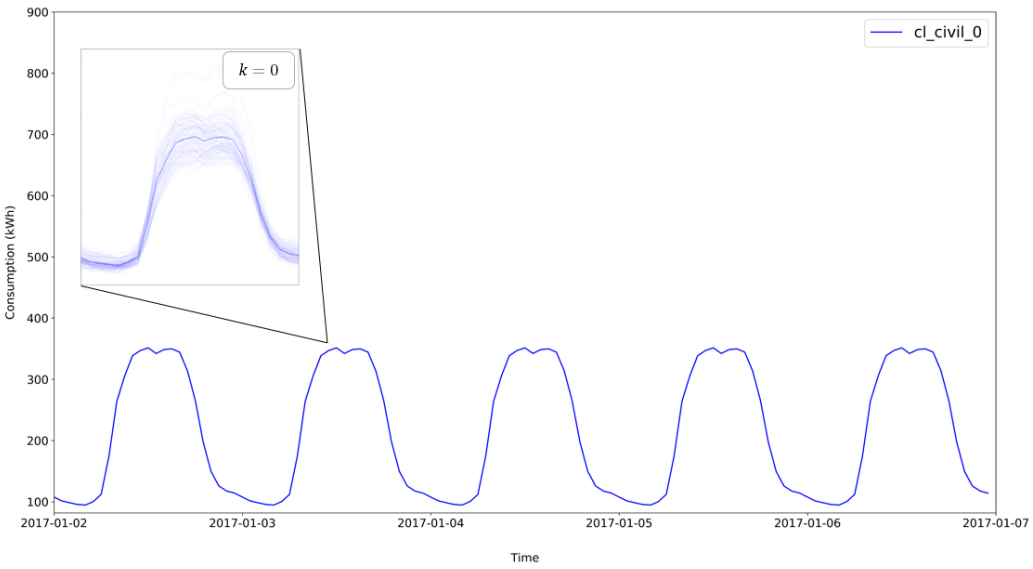


Figure 4.7 - Civil Building cluster average feature from k=0 (in 3.2.2 daily analysis - Figure 3.7)

Therefore, with this process the same number of features as clusters per building were generated, specifically, three for Civil, Central, and North tower and four for South tower. To identify each feature the correspondent cluster's number (k) was used as feature name suffix, e.g. for cluster k = 0 of Civil building the feature name is cl_civil_0.

Afterwards, the second and last group of generated features, was defined considering the common use of lagged features in time series, referred in Table 2.2 of section 2.2.2. This new group of features was based in the autocorrelation applied to each building energy consumption. The autocorrelation was obtained using Pearson's correlation coefficient between a chosen number of previous hours and the actual hour. From the latter calculations, it was concluded that all the study cases have an identical behaviour, therefore, Civil building autocorrelation of the previous seven days is used as role model, Figure 4.8.

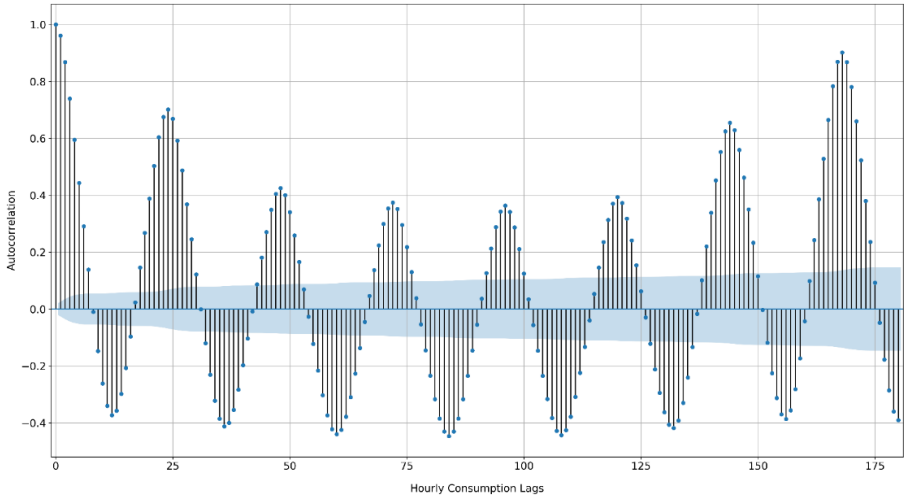


Figure 4.8 - Autocorrelation of Civil building hourly energy consumption

From Figure 4.8, it is visible that the three most autocorrelated periods, take place first at the previous hour, then a week before, and lastly at the previous day. These three periods were used to generate lagged features for each building. As a consequence of that, the first week of 2014 was lost. The new features were named using the lagged period as suffix, e.g. $civil\_lag1hour$ for civil building previous hour.

To complement each of the chosen periods, it was also created three other features, that provide the minimum, the maximum and the average of the three hours prior to each period, through the use of a rolling window technique. These new features were chosen not only due to the high autocorrelation of the last 3 hours, but also to provide to the lagged features a certain continuity. The reason why it was selected a rolling window technique and not directly used the three hours prior the periods is based on absence of losing data. For the rolling window features name it was added a suffix to the lagged feature that characterizes the period that was selected from, e.g. for civil building 1 hour lagged rolling window maximum, $civil\_lag1hour\_rollmax$.

Nonetheless, each data set created in 4.1 was split per building and time horizon, based on the set of features that were available to use for each of the different forecasting horizons (an hour, a day, and a week) represented in Table 4.4. Where (a) may be replaced by each building name. From this table, there was a total of 27 features for Civil, Central, and North tower buildings and 28 features for South tower building.

Table 4.4 - Set of feature per building (a) and forecasting horizon for each of the data sets

Data set (dt_01, dt_02, and dt_03)					
Based on		**Features**	**1 hour horizon prediction**	**1 day horizon prediction**	**1 week horizon prediction**
Time and Calendar		t_hour	✔	✔	✔
		$t\_dayofweek$	✔	✔	✔
		t_month	✔	✔	✔
		$s\_workday$	✔	✔	✔
		s_epoch	✔	✔	✔
Energy Consumption	Cluster Average	$cl\_(a)\_0$	✔	✔	✔
		$cl\_(a)\_1$	✔	✔	✔
		$cl\_(a)\_2$	✔	✔	✔
		$cl\_(a)\_3$ - just for South tower	✔	✔	✔
	Lagged Features	$(a)\_lag\_1hour$	✔	-	-
		$(a)\_lag\_1hour\_rollmin$	✔	-	-
		$(a)\_lag\_1hour\_rollmax$	✔	-	-
		$(a)\_lag\_1day$	✔	✔	-
		$(a)\_lag\_1day\_rollmin$	✔	✔	-
		$(a)\_lag\_1day\_rollmax$	✔	✔	-
		$(a)\_lag\_1day\_rollmean$	✔	✔	-
		$(a)\_lag\_1week$	✔	✔	✔
		$(a)\_lag\_1week\_rollmin$	✔	✔	✔
		$(a)\_lag\_1week\_rollmax$	✔	✔	✔
		$(a)\_lag\_1week\_rollmean$	✔	✔	✔
Weather Conditions Data (WCD)		wt_temp	✔	✔	✔
		$wt\_meansolarrad$	✔	✔	✔
		wt_hr	✔	✔	✔
		$wt\_mean\_windspd$	✔	✔	✔
		$wt\_max\_windgust$	✔	✔	✔
		wt_mean_pres	✔	✔	✔
		$wt\_mean\_solarrad$	✔	✔	✔
		wt_rain_day	✔	✔	✔

4.3 Models Selection

After the previous data treatment and features generation, the prerequisites were met to proceed to the machine learning models selection and posterior evaluation of the different data sets created in 4.2.

The machine learning models used, in this study, were four models: ANN(MLP), SVM, RF, and XGB. The first three models were given by scikit-learn [70] and the last one by XGBoost [71] python packages. In addition, ANN(MLP) and SVM were chosen taking into account their substance use in the literature reviewed (2.2). The other two models, RF and XGB, despite their insignificant usage when compared with the first two models in forecasting, they are emerging in the machine learning field, showing in recent studies, [61] and [62], better performance than the first two models.

Furthermore, to evaluate and compared the chosen models, each created data set, was split into two subsets: the training and the testing set. The training set was responsible for the learning process of every model and the posterior comparison of the diverse strategies employed during this work. It consists of the first two years of data, specifically, 2014 and 2017. The rest of the data, 2018, was used only to test the model's ability to forecast, using the error metrics mention in 2.4.

4.3.1 Data Normalization

It is also worth to mention the need or not of data normalization, also known as feature scaling. This pre-processing step is used depending on the machine learning model used. It is known to be beneficial to some of machine learning models that are comparing features with different scales, such as k-means clustering, previous used in 3.2.2. In terms of the above mention models used for forecasting, the only ones that required data normalization to yield better predictions are the ones that are not based in decision trees, specifically ANN(MLP) and SVM.

There is no obvious answer to which type of normalization should be used. One often used is defined by equation (4.2), that makes every feature from the training data to be confined in values from 0 to 1.

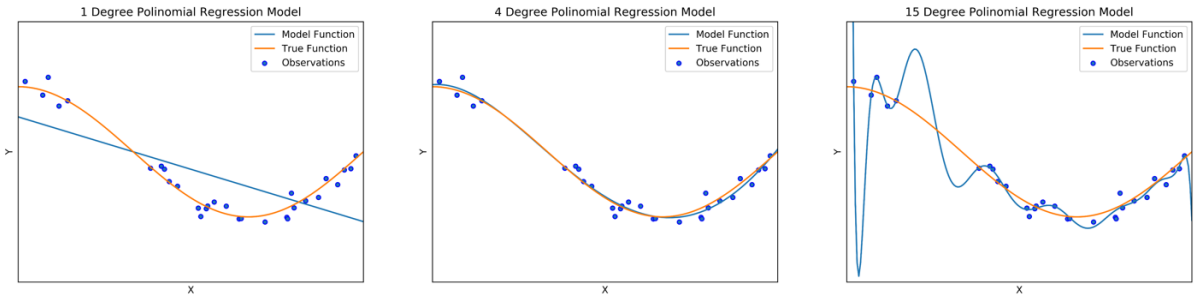
$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4.2}$$

Where, X refers to each feature vector, X_{min} and X_{max} the minimum and the maximum encounter in the feature vector, respectively, and $X_{normalized}$ denotes the outcome normalized feature vector.

4.3.2 Cross-Validation (CV)

One of the main objectives when training a ML model is its generalization performance in the unseen data, known as testing set. In the opposition to the generalization, phenomenons of under and over fitting may occur when the model is trained, leading to poor performance on the testing set. In the case of under fitting the model is not suitable for the complex nature of the addressed problem being not capable of neither modelling the training set nor generalize to unseen data, Figure 4.9 (a). On the other hand, over fitting occurs when the model learns the detail and noise in the training set to the extent that it negatively impacts the performance of the model in the testing set, Figure 4.9 (b).

In the circumstances of this work, since the chosen models are robust enough to handle the non-linearity found in the data set, the only adverse scenario that could be found is over fitting. To overcome this, a cross-validation technique (CV) in the training data was used.



(a) Under fitting (b) Good fitting (a) Over fitting

Figure 4.9 - Example of under (a), good (b) and over (c) fitting for a polynomial regression model

When handling with time dependent data, the standard K-fold CV technique is not suitable, once it naively ignores their inherent sequential nature of time. To circumvent this problem a specific time series technique was used, denoted in this dissertation as ts-CV. This technique splits the training data set, in K folds and uses the first K fold as training set and the $K + 1$ as the testing set. The process is repeated successively using each time, the past values to predict future ones, guaranteeing the time dependency needed, as illustrated in Figure 4.10.

Training	Test	-	-	-
Training	Training	Test	-	-
Training	Training	Training	Test	-
Training	Training	Training	Training	Test

Figure 4.10 - Example of a CV technique adapted for time dependent data

Nonetheless, in this study, the ts-CV split the training set into 24 folds, despite the intrinsic computational expenditure. This ensured that the models for each decision making trained the data month by month, allowing the contribution of each month in the average error.

4.4 Feature Selection

The objective of this section is to select the most relevant created features, before modeling the data sets of each building and forecasting horizon (an hour, a day, and a week). This procedure, when properly applied, its known to improve the models, in terms of over fitting, accuracy, and by reducing the training time. The former occurs, since with less redundant features, the probability of the models to make decisions based on noisy or misleading data is lower and by using less data the algorithms reduce their complexity and train faster.

There are three general classes of feature selection: filter, wrapper, and embedded methods. Filter methods apply statistical measures to assign a score to each feature, the methods are usually univariate and consider each feature independently. The wrapper methods consider the selection of a set of features as a search problem, where different combinations of features are prepared, evaluated and compared with other combinations. To evaluate those combinations a machine learning model is used

to score each group of features based on an error metric prior established. Lastly, the embedded methods, learn which features best contribute to the accuracy of the model while the model is being created, which is often done by regularization methods. A detailed explanation of each class and its advantages, is given in [67].

In this dissertation it was firstly used a filter method by Pearson correlation, to elect the most relevant WCD features, and afterwards, with the remaining features, a wrapper method, named as recursive feature elimination (RFE), was employed using XGB model, Figure 4.11.

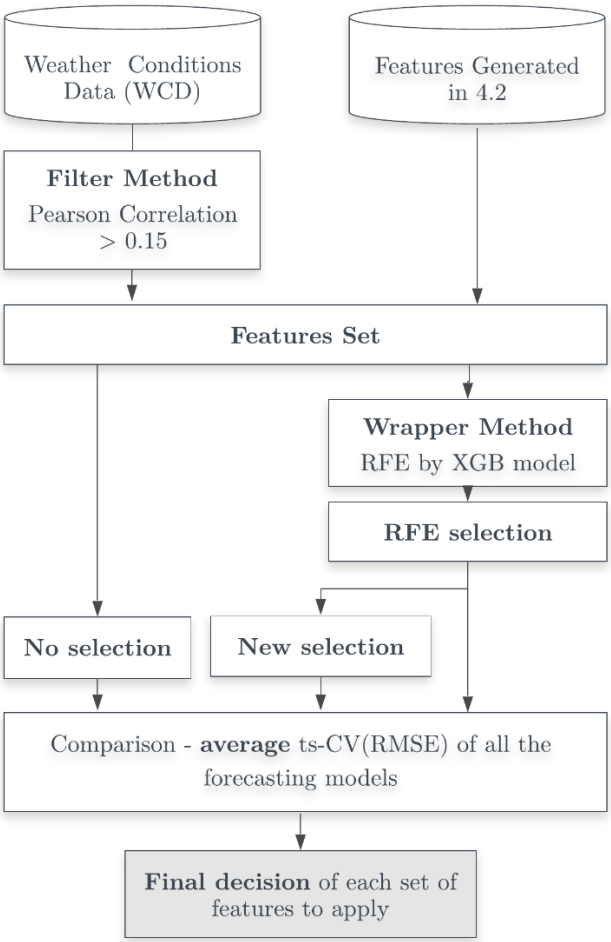


Figure 4.11 - Procedures diagram of feature selection

Filter method - Pearson Correlation

When Pearson correlation between each of WCD features and the corresponding buildings was implemented, it was found that half of the available features had a correlation below 0.15 with every building. That features were automatically removed, since their lack of information about each building dynamic behavior, as it can be seen in Figure 4.12. As a result, the only useful features were wt_temp, wt_tmpap, wt_hr, and $wt\_mean\_solarrad$.

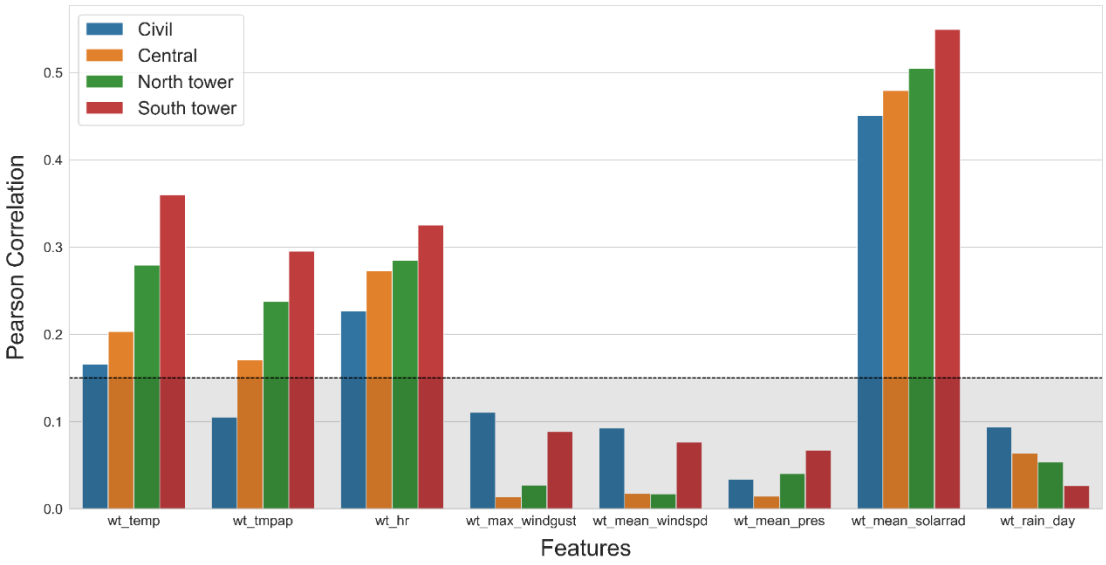


Figure 4.12 - Pearson correlation between WCD features and each building

Furthermore, to check the correlation between each of the selected WCD features, a heatmap was done and shown in Figure 4.13. The obvious was observed in the correlation between the almost similar features: wt_temp and wt_tmap, that accordingly to Table 4.1, represent the temperature and the apparent temperature, respectively. With that in mind, the first was elected since it presents higher correlation score with every building, Figure 4.12.

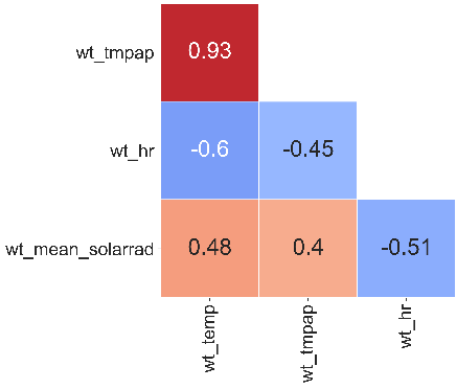


Figure 4.13 - Heatmap made by Pearson correlation score between the chosen weather conditions features

Finally, the set of WCD features were enclose with the features generated in 4.2, so the wrapper method could be performed**.**

Wrapper method - RFE

The RFE as a wrapper method, uses a given external model, that assigns weights to each of the features. First, the model is trained on the initial set of features and to each of them is attached a weight that ranks their importance, through an attribute embedded in the model. Afterwards, the least important feature is eliminated from the current data set. That process is recursively repeated on the remaining features, until the data set is reduced to one feature. As a result, the method supplies the cross-validation score for each of the combinations and choses the best set of features. Among the models selected to perform the forecasting, only two models, specifically RF and XGB had the embedded

attribute required to implement the RFE method. However, due to the computation expenditure of this process it was only performed by XGB model due to its computational speed characteristics.

To ensure the consistency of the selection performed by the RFE method, each set of features selected per time horizon and building was visualized and discussed. After that, some features that were considered as important based on the knowledge acquired in 3.2.2 energy analysis were added to the RFE method selection, creating a new selection of features.

Lastly, the forecasting models chosen in 4.3 were used to calculate the average of the ts-CV(CV(RMSE)) error to compare: the inexistence of selection process, the RFE method selection, and the new selection. Taking into account that error metric and the posterior hyperparameter optimization, a set of features was selected to feed each of the models per building and time horizon.

4.5 Hyperparameter Optimization

In machine learning algorithms there are two types of parameters: the model parameters and the hyperparameters. The model parameters are defined during the learning process, meaning that the model adapts each parameter with the objective of fitting the best way possible a given data set, e.g. the weights addressed to each neuron in an ANN. On contrary, the hyperparameters cannot be learned within the model directly, are unchangeable since the moment they are pre-defined, and play a crucial role in the training stage and posterior model predictions, e.g. the number of hidden layers and learning rate in an ANN.

Therefore, since the latter parameters may be defined beforehand, is of great importance to choose then wisely considering the type of data that feeds the model. In addition, as it is known, a major part of machine learning algorithms possesses a large variety of hyperparameters, which may be translated in a misleading and expensive work when selected manually. To address that issue, hyperparameter searching techniques are usually applied.

Hyperparameter searching techniques when applied by trial and error, such as grid search and random search, may lead to an inefficient and time-consuming process, since they roam the given space of available hyperparameters values in an isolated way without paying attention to past results. To ease this process, techniques such as evolutionary algorithms and Bayesian optimization may be employed, as it was in [48] and [54], respectively. In this work a Bayesian optimization grounded by Gaussian processes was used. A detailed explanation of the later technique and its advantages may be found in [68]. To apply this technique a python package, named as Scikit-Optimize [69] was used.

For the Bayesian optimization an hyperparameter search space for each model was pre-defined. Since the models chosen for forecasting have different characteristics, their hyperparameters and consequent spaces were quite different. A detailed information about the hyperparameters search space for each model may be seen in Table 4.5.

Table 4.5 - Hyperparameter search space for each of the models

Models	Search Space
ANN(MLP)	
Hidden layers sizes	(100); (100, 20); (100,20,20); (100,20,20,20); (40, 100, 40, 40); (40, 100, 40); (20, 100, 20, 20); (20, 20, 100, 20);
Activation function	ReLU, tanh
Learning rate	Adaptive, Invascalling
Batch size	24; 48; 168; auto
SVM	
C	Logarithmic uniform distribution [0.01; 100]
gamma	Logarithmic uniform distribution [0.001; 10]
epsilon	auto; scale
RF	
Nº of estimators	Integers values from 100 to 1000
Max. features	auto; sqrt
Min. Samples split	Integers values from 2 to 40
Min. Samples leaf	Integers values from 1 to 20
XGB	
Nº of estimators	Integers values from 100 to 600
Learning rate	Logarithmic uniform distribution [0.001; 1]
Max. depth	Integers values from 1 to 30
Subsample	Logarithmic uniform distribution [0.1; 1]

It is worth to mention, that the hyperparameters chosen to be search, were based on the default values that each model presents in the correspondent python package. A detailed information of each hyperparameter characteristic is given in [70] for ANN(MLP), SVM, and RF models, and in [71] for XGB model.

Chapter 5

Results

In this Chapter, the models chosen in 4.3, were compared and used for decision making of the diverse strategies adopted during the methodology. The evaluation of that strategies was performed in the training set, using the average error of the four models. The error metric applied for this comparison was the CV(RMSE) error (2.4), calculated by the ts-CV, referred in 4.3.2.

The diverse strategies, specifically, the data imputation study and posterior data sets generation in 4.1, the features selection performed in 4.4, and the hyperparameter optimization of 4.5, were divided by sections to ease the visualization and consequent decision. After that, the last section reveals the forecasting results of each developed model for every building (Civil, Central, North tower, and South tower) and time horizon (an hour, a day, and a week). To evaluate the forecasting results the error metrics referred in 2.4 were used.

5.1 Data Imputation Study

As it was mentioned in 4.1, two multiple imputation algorithms (MICE and MF) were studied with the intend of filling the gaps of the original data set. This study was performed in the data set of 2014 for 10 cycles with randomly created artificial gaps. The results for ECD and WCD may be seen in Table 5.1 and Table 5.2, respectively.

For the ECD the results show a clear dominance in terms of accuracy for MF algorithm. Based on that, this algorithm was chosen for the posterior data sets generation. For further details of both algorithms imputations capabilities, Civil, North tower, and South tower last cycle imputations may be seen in Figure 5.1, Figure 5.2, and Figure 5.3, respectively. The Central building is not displayed since only one missing value was imputed per cycle, according to its target defined in Table 4.1.

Table 5.1 - Mean Error Evaluation of 10 cycles for MF and MICE algorithms in ECD, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units

Multiple Imputation model	MICE			MF		
Error Metrics (%)	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE
$civil$	14.816	14.815	18.097	**2.464**	2.463	3.438
$central$	32.882	27.776	45.814	**12.321**	8.981	14.965
$north\_tower$	54.383	42.506	34.338	**20.132**	11.969	10.661
$south\_tower$	38.649	23.368	43.276	**16.856**	8.358	16.474

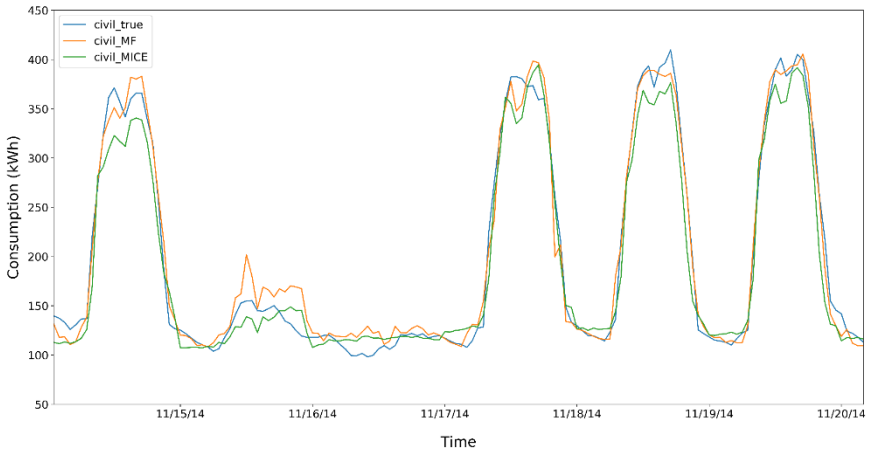


Figure 5.1 - Civil building last cycle imputations with MF and MICE (148 values)

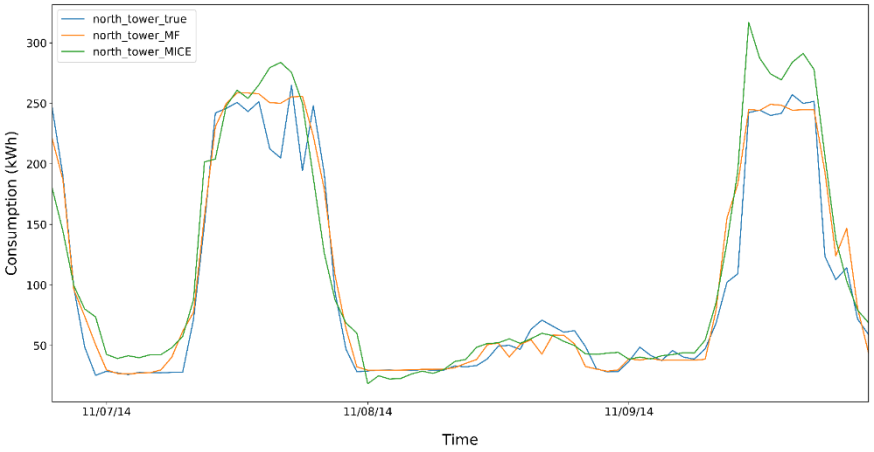


Figure 5.2 - North tower last cycle imputations with MF and MICE (78 values)

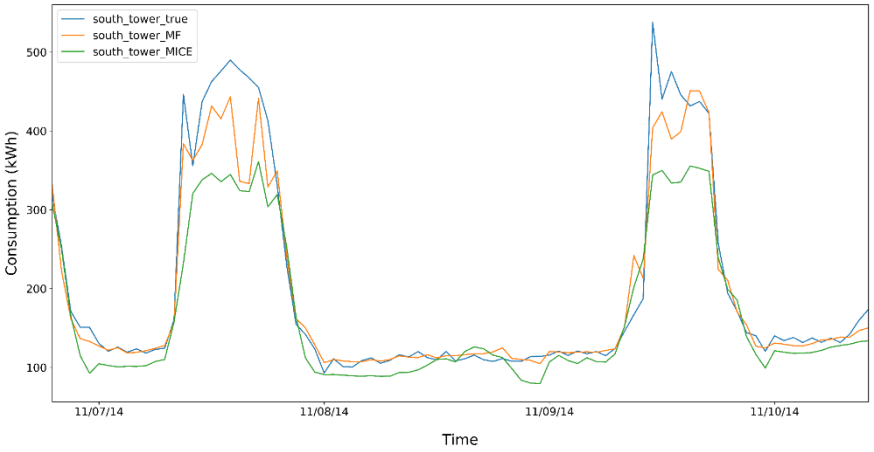


Figure 5.3 - South tower last cycle imputations with MF and MICE (88 values)

For the WCD, the scenario was quite different. The errors values although slightly better for MF algorithm, were too high to be used in the posterior data sets generation. This performance difference between WCD and ECD can be explained by the fact that the WCD has a larger gap of missing values occurring in simultaneous for all the features to impute. Moreover, the WCD features are not as similar as each building consumption in the ECD.

Due to the poor performance found in WCD imputation, the HMM method referred in 4.1 was employed to fill the missing values for dt_03.

Table 5.2 - Mean Error Evaluation of 10 cycles for MF and MICE algorithms in WCD, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in feature units (Table 4.1)

Multiple Imputation model	MICE			MF		
Error Metrics (%)	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE
wt_temp	53.040	46.124	7.109	**26.344**	22.929	3.551
wt_tmpap	76.106	74.067	8.069	**38.418**	37.720	4.082
wt_hr	50.559	51.309	24.879	**26.281**	26.797	12.723
$wt\_mean\_windspd$	88.447	10317.186	5.544	**44.299**	5231.062	2.766
$wt\_max\_windgust$	98.556	8105.362	3.393	**49.377**	4093.348	1.695
wt_mean_pres	1.460	1.120	11.352	**0.791**	0.614	6.231
$wt\_mean\_solarrad$	231.381	208341.775	320.780	**88.130**	1096.834	97.986
wt_rain_day	1013.131	182924.141	3.737	**630.617**	97027.176	2.028

5.2 Data sets analysis

In this analysis, the data sets generated in 4.1 were compared. To procced with it the average error of the four models chosen in 4.3 (ANN(MLP), SVM, RF, and XGB) in an hour, a day, and a week horizon is shown in Table 5.3.

Table 5.3 - Average ts-CV(CV(RMSE)) error of the four models for each data set generated in 4.1, displayed by building and time horizon

Time horizon	1 hour			1 day			1 week		
Data sets	dt_01	dt_02	dt_03	dt_01	dt_02	dt_03	dt_01	dt_02	dt_03
Civil	13.388	13.800	**13.383**	37.543	**37.490**	38.975	**39.248**	39.605	40.453
Central	16.358	16.430	**15.738**	39.693	40.185	**39.433**	44.438	45.173	**42.383**
North tower	23.970	**23.685**	23.873	45.523	45.735	**42.900**	48.333	48.160	**44.805**
South tower	24.755	**24.690**	24.895	39.170	39.080	**38.428**	41.470	41.685	**40.110**
Average	19.618	19.651	**19.472**	40.482	40.623	**39.934**	43.372	43.656	**41.938**

In general, from Table 5.3, it is possible to conclude that despite the slight error variations between the data sets, the one with the most accurate results was dt_03. This might be explained by the absence of data lost, as mentioned in Table 4.3, which gives the models the possibility to improve their predictions since there is more data to learn from.

Additionally, the use of dt_03 from this point forward allowed the prediction of every hour of 2018 year, which was one of the reasons why the data imputation study in 4.1 was conducted in first place.

5.3 Feature Selection Analysis

From the RFE method, mentioned in 4.4, to each building and time horizon a set of features was selected. An example of this method application for North tower building for each of the horizons data sets is shown in Figure 5.4.

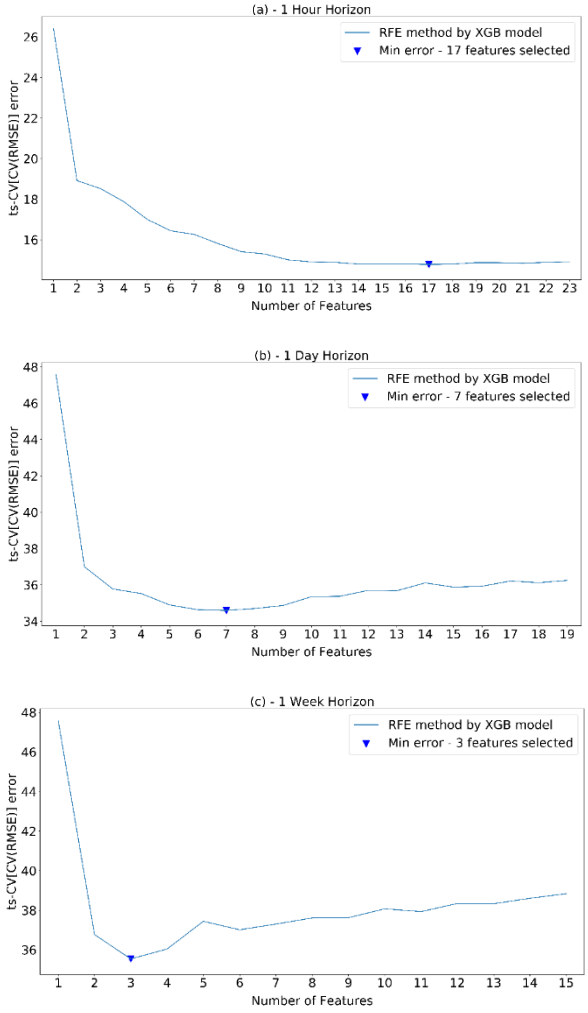


Figure 5.4 - North tower building RFE method application by XGB, for an hour (a), a day (b), and a week (c) horizon data sets

As it is possible to see from Figure 5.4 the features that were selected for each of the time horizon tend to decrease in ratio when the horizon of prediction increases. In this particular case, the ratios were 0.76, 0.33, and 0.14, for an hour, a day, and a week, respectively.

Apart from the fact that in an hour horizon data set there are more features than in a day or a week horizon, several features that were considered as important in the first data set were neglected in the other two. With that in mind, the creation of a new set of features supported by the RFE method selection and the knowledge acquired from each building energy analysis in 3.2.2 was performed. To ease the visualization of that the Table 5.4 was created. Where, the black check marks (✔) indicate the features that were selected by the RFE method using the XGB model and the red check marks (✔) represent the features that were added to that selection to create the new set of features, referred in this dissertation as new selection.

Table 5.4 - Features selected by the RFE method (✔) and features that were added to that selection (✔) for each building and time horizon. Where H, D, and W denotes an hour, a day and a week horizon, respectively.

Buildings (a)	Civil			Central			North tower			South tower			Usage
Forecasting horizon data set	H	D	W	H	D	W	H	D	W	H	D	W	(%)
t_hour	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	33
$t\_dayofweek$		✔		✔	✔		✔	✔			✔		50
t_month		✔									✔		17
$s\_workday$	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	**100**
s_epoch	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	33
$cl\_(a)\_0$	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	**100**
$cl\_(a)\_1$	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	67
$cl\_(a)\_2$	✔						✔	✔	✔	✔	✔	✔	25
$cl\_(a)\_3$	-	-	-	-	-	-	-	-	-	✔	✔	✔	33
$(a)\_lag\_1hour$	✔	-	-	✔	-	-	✔	-	-	✔	-	-	**100**
$(a)\_lag\_1hour\_rollmin$	✔	-	-	✔	-	-	✔	-	-	✔	-	-	**100**
$(a)\_lag\_1hour\_rollmax$		-	-	✔	-	-	✔	-	-	✔	-	-	75
$(a)\_lag\_1hour\_rollmean$		-	-	✔	-	-		-	-	✔	-	-	50
$(a)\_lag\_1day$	✔	✔	-	✔	✔	-	✔	✔	-	✔	✔	-	**100**
$(a)\_lag\_1day\_rollmin$	✔	✔	-	✔	✔	-	✔	✔	-	✔	✔	-	86
$(a)\_lag\_1day\_rollmax$		✔	-	✔	✔	-	✔	✔	-		✔	-	50
$(a)\_lag\_1day\_rollmean$		✔	-			-	✔	✔	-	✔	✔	-	38
$(a)\_lag\_1week$	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	**100**
$(a)\_lag\_1week\_rollmin$	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	50
$(a)\_lag\_1week\_rollmax$		✔	✔	✔	✔	✔	✔	✔	✔		✔	✔	42
$(a)\_lag\_1week\_rollmean$		✔	✔			✔	✔	✔	✔		✔	✔	25
wt_temp	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	33
$wt\_mean\_solarrad$	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	42
wt_hr	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	17

Before presenting the reasons why the features checked with the red mark were added, it is important to acknowledge that three type of features were always included in the RFE method selection. Specifically, the $s\_workday$, the $cl\_(a)\_0$, and each different lagged feature when available ($(a)\_lag\_1hour$, $(a)\_lag\_1day$, and $(a)\_lag\_1week$). For further details about the features elected by this method, a column with the percentage of usage was supplied in Table 5.4.

Furthermore, the reasons that support the addition of each feature to the RFE method selection were the following:

- t_hour - due to the temporal granularity of the data and since it was always selected by the RFE method for the hour horizon data set of each building;
- s_epoch - mainly due to the support that provides to the $s\_workday$ feature to address the stochastic behaviour of the daily inhabitants of each building, in terms of the different types of occupancy occurring during the different periods of the school calendar;
- For the cluster average features, e.g. $cl\_(a)\_0$, the only ones that were added were the ones that exhibit the most different average consumption patterns, specifically: $cl\_(a)\_0$ and $cl\_(a)\_1$ for Civil and Central building; $cl\_(a)\_0$, $cl\_(a)\_1$, and $cl\_(a)\_2$ for North tower

building; and $cl\_(a)\_0$, $cl\_(a)\_1$, $cl\_(a)\_2$, and $cl\_(a)\_3$ for South tower building. As it may be seen in daily consumption analysis of 3.2.2;

- Lastly, the weather conditions features were added in every data set, firstly because of their good correlation with each building, shown in Figure 4.12, and secondly, since they are the only features that contain information about the exact moment of prediction.

Afterwards, with the new selection defined, similar to what was done in 5.2, the average of the four models was employed to compare the inexistence of selection, the RFE method selection and the new selection. The results of that comparison for each building and time horizon may be viewed in Table 5.5.

Based on the table below, it is possible to conclude that for every building and time horizon the inexistence of selection had always achieved higher average error values than any of the selection performed, guaranteeing with that the importance of this type of analysis.

Table 5.5 - Average ts-CV(CV(RMSE)) error of the four models for each set of feature used (No selection, RFE method selection, and the new selection) by building and time horizon.

Time horizon	1 hour			1 day			1 week		
Selection	No sel.	RFE	New sel.	No sel.	RFE	New sel.	No sel.	RFE	New sel.
Civil	11.799	**11.350**	11.459	34.524	34.038	**31.074**	37.095	**33.231**	36.180
Central	14.765	14.385	**14.316**	37.172	34.302	**34.030**	41.018	**36.823**	37.125
North tower	19.546	**17.102**	17.399	40.606	38.447	**38.042**	43.570	**40.010**	43.193
South tower	21.939	**19.316**	19.633	34.937	**34.626**	34.853	37.577	**35.474**	37.313

Moreover, it is important to state that the models which performed these average errors did not suffer any kind of optimization process. That being said, although for some cases there were small errors deviations that leaned towards the RFE method selection, this selection was not used. The reason for that is based on the restriction of the number of features that the models may learn from when optimized. An example of that restriction occurs in the week horizon features selection for three of the four buildings, where just $s\_workday$, $cl\_(a)\_0$, and $(a)\_lag\_1week$ features were selected to performed the prediction.

So, in order to not restrict the models learning process, the new selection of features was chosen to feed each model for the hyperparameter optimization, despite the errors shown in Table 5.5.

5.4 Bayesian Optimization

To optimize every model for each building and time horizon data set a Bayesian optimization was performed during 30 iterations using the hyperparameter search space define in Table 4.5. This number of iterations was chosen due to the implicit computational cost of each model optimization, since it was done for the 48 models. The hyperparameters that were chosen in that search are given in Table 5.6.

Table 5.6 - Hyperparameter selection to each building and forecasting horizon using Bayesian optimization, where H, D, and W represents the hour, the day and the week horizon models.

Buildings	Civil			Central			North tower			South tower		
Time horizon	H	D	W	H	D	W	H	D	W	H	D	W
ANN(MLP)												
Hidden layers	(100,20)	(100,20,20)	(100,20,20)	(100,20)	(100,20)	(40,100,40,40)	(40,100,40)	(20,100,20,20)	(100,20)	(100)	(100)	(100,20)
Activation function	ReLU	tanh	ReLU	ReLU	ReLU	ReLU	ReLU	tanh	ReLU	ReLU	tanh	ReLU
Learning rate	Adapt.	Adapt.	Invscal.	Adapt.	Adapt.	Adapt.	Invscal.	Invscal.	Invscal.	Invscal.	Invscal.	Invscal.
Batch size	24	168	auto	24	auto	24	48	24	48	168	48	168
SVR												
C	13.091	0.039	0.018	100.000	0.643	4.612	6.500	0.658	0.104	15.789	3.889	1.138
Gamma	auto	scale	scale	auto	auto	scale	auto	auto	scale	auto	auto	auto
Epsilon	0.0095	0.0020	0.0012	0.0052	0.0019	0.0519	0.0460	0.0201	0.0131	0.0388	0.0046	0.0096
RF												
Estimators	100	100	100	1000	1000	1000	158	842	1000	1000	655	100
Max. features	auto	sqrt	sqrt	auto	sqrt	sqrt	auto	sqrt	sqrt	auto	sqrt	sqrt
Min. samples split	14	2	2	6	20	2	6	20	2	14	2	2
Min. samples leaf	4	20	20	5	20	20	3	16	20	5	20	18
XGB												
Estimators	580	320	540	600	358	500	420	483	155	240	460	400
Learning rate	0.024	0.029	0.007	0.025	0.049	0.023	0.017	0.019	0.100	0.085	0.031	0.015
Max. depth	7	15	6	26	239	2	24	105	2	7	23	4
Subsample	0.964	0.196	0.334	0.441	0.343	0.740	0.368	0.339	1	0.869	0.148	0.741

5.5 Forecasting

As a result of all the previous studies completion, this section has the objective of showing the forecasting capabilities of the four machine learning models used for the three time horizons and the four buildings studied (4 x 3 x 4 = 48 models), in the test set (year of 2018).

To better define their capabilities, this section was divided into four subsections where each one addresses one building at the time. For each of the subsections every time horizon was presented with the correspondent forecasting models' errors, referred in 2.4. For every time horizon the best forecasting model was selected, and its monthly errors were shown. Based on the monthly errors, two weeks, one from the best and another one from the worst month, were graphically visualized against the true energy consumption. To ensure this work consistency the CV(RMSE) error metric was once again used to elect the best forecasting model per time horizon and building.

Furthermore, the error metrics for each type of day were also displayed for the best forecasting models prior elected. The type of days included in this analysis were workdays, holidays, summer break days and weekends. In addition, to somehow show models potentialities an atypical week is shown for each of the buildings.

5.5.1 Civil Building

In Civil building, for every time horizon the forecasting results did not exceed an CV(RMSE) error of 16.75%, which according to Table 5.7 was achieved in a week horizon prediction by XGB model. According to that, it is noticeable an increasing error tendency when the models attempt to predict in greater horizons, e.g. the RF model obtained CV(RMSE) errors of 7.14%, 12.18%, and 14.29% for an hour, a day, and a week horizon predictions, respectively. It may also be concluded that all the models tend to adapt easily when predicting an hour a-head consumption, achieving similar prediction accuracies, on contrary to what occurs in a day and a week a-head predictions, where the best model was easily selected.

Table 5.7 - Annual results for Civil building forecast for an hour, a day, and a week horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units

Models	1 hour			1 day			1 week		
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE
MLP	6.93	5.04	8.90	13.42	10.94	17.33	16.65	12.50	22.11
SVM	**6.32**	**4.22**	**7.86**	**11.09**	**7.69**	**13.97**	**14.00**	**10.10**	**18.10**
RF	7.14	4.44	8.42	12.18	8.63	15.06	14.29	10.75	18.61
XGB	6.52	4.22	7.92	14.30	11.00	18.34	16.75	12.98	22.26

Furthermore, from all the models used, the SVM model achieved the best annual forecasting results for all the horizon predictions, in the three use error metrics. Consequently, this model was used to check the monthly predictions accuracy for each of the horizons, Table 5.8.

Table 5.8 - Civil building monthly forecast results of the best models selection for each time horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units

Months	1 hour - SVR			1 day - SVR			1 week - SVR		
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE
January	6.93	4.45	8.48	10.85	7.54	13.34	11.29	8.15	14.92
February	6.38	3.91	7.89	12.16	8.18	15.85	18.01	9.94	21.68
March	6.16	4.05	7.92	10.39	7.37	13.88	12.91	9.32	17.93
April	6.35	4.30	8.04	11.47	7.79	14.93	14.80	10.52	18.83
May	5.93	4.10	7.55	11.10	7.25	14.04	11.53	8.29	15.63
June	5.70	4.29	8.01	9.83	7.68	14.11	13.76	10.87	19.86
July	5.81	4.12	7.57	**8.12**	**5.85**	**10.73**	**9.68**	**7.77**	**13.74**
August	8.74	4.52	**6.47**	14.37	9.35	12.07	17.41	14.77	17.17
September	7.30	4.41	9.20	12.77	6.93	15.84	18.97	10.79	25.25
October	**5.52**	**3.86**	7.81	8.77	6.70	12.89	10.19	7.85	15.09
November	5.68	4.16	8.10	10.83	8.47	15.34	11.40	9.51	17.11
December	5.95	4.49	7.48	12.76	9.18	14.82	15.69	13.08	19.95

Based on Table 5.8, the months that yielded better results were October for an hour horizon, and July for a day and a week horizon. This might be explained by the small daily consumption variations felt in those months, making the correspondent real consumption ease to predict when compared to other months. On the other hand, the worst case scenario was found in August for an hour and a day horizon prediction, and in September for a week horizon prediction. The unpredictability of these type of months derivates from particular events that somehow change abruptly the energy consumption of the building, such as the two weeks of summer break in August and the beginning of the first semester in September, regardless the used of the $s\_workday$ and s_epoch features in attempt to recognize this type of behaviours. After all, two weeks of July and August, were chosen to be visualized, as an example of the best and worst case scenarios for the all the forecasting horizons, respectively.

In Figure 5.5, as it was expected, the hour horizon model outperforms the other time horizon models, even though July did not represent the highest accuracy. In general, all the time horizons achieved accurate results, showing slightly deviations from the true energy consumption, in more evident in weekdays than in weekends.

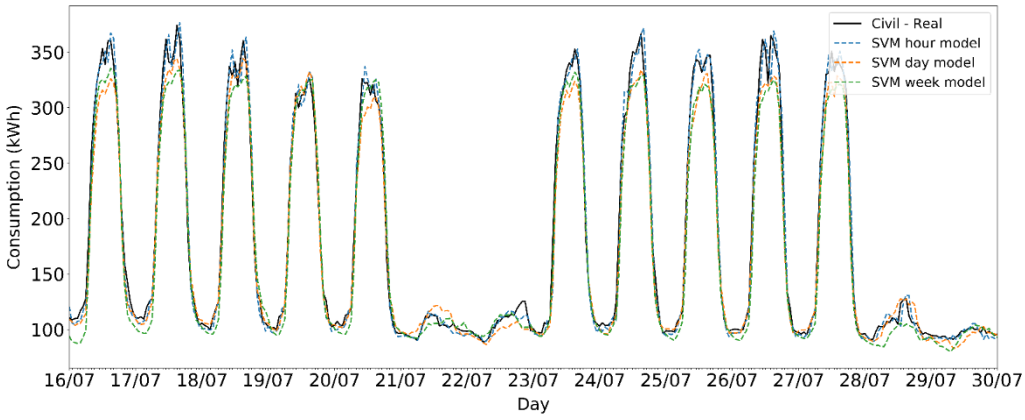


Figure 5.5 - Forecasting of two weeks of July for each time horizon best model

For the two weeks of summer break shown in Figure 5.6, the predictions were not so good. However, for an hour horizon the predictions followed the true consumption almost perfectly, being the biggest deviation found in the first day of vacations, with an unexpected peak of energy consumption. For a day horizon the forecasted consumption was able to follow the building base energy consumption decrease along the first week of vacations, although during the day the predictions tend to replicate the previous day, an example of this may be seen from the fifteenth to the sixteenth day of the month. Lastly, the predictions performed in a week horizon were not able to adapt so easily to the building base energy consumption being barely above the true consumption during the night period.

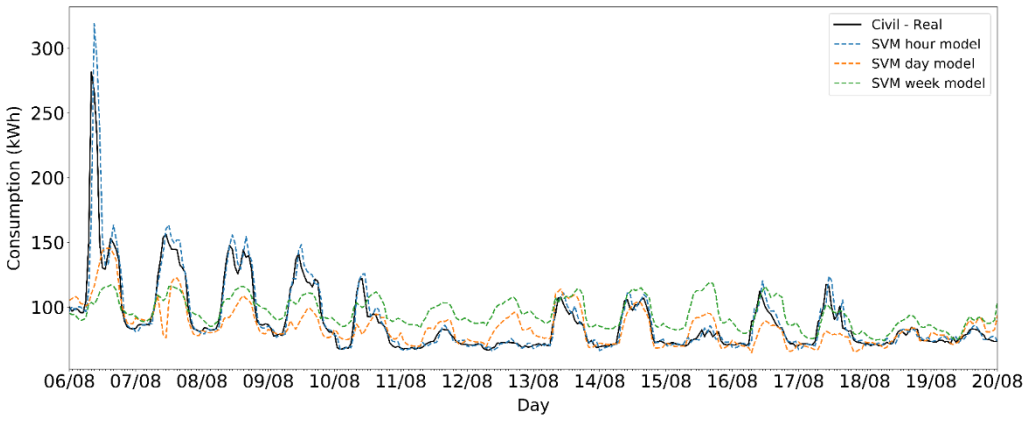


Figure 5.6 - Civil building forecasting of the two weeks summer break for each time horizon best model

From the day type analysis results, in Table 5.9, it may be seen that the workdays were in most of the time horizons the easiest type of day to predict, which means that in 65% of the days in the three forecasting horizons the MAPE values were below 10%. Right after those type of days, weekends that account with 28% of all the days, obtained the second best predictions, displaying MAPE values below 11%. That being said, it is possible to conclude that 93% of the days (workdays and weekends) were successfully forecasted in the three time horizons when compared to the remaining 7% of the days that included holidays and the two weeks of summer break, previously shown in Figure 5.6.

Table 5.9 - Day type results for Civil building best models forecast of each time horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units

Day type	1 hour - SVM			1 day - SVM			1 week - SVM			Day (%)
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	
Workdays	6.01	4.45	9.70	**9.89**	**6.93**	15.72	**12.79**	**9.16**	20.59	65
Weekends	**4.89**	**3.35**	**3.79**	12.52	8.07	**9.15**	16.00	10.95	**12.19**	28
Holidays	9.24	6.46	7.99	26.24	17.59	22.02	28.17	20.86	25.01	4
Summer break	11.52	4.47	4.91	22.77	10.68	11.74	26.16	18.16	17.14	3

5.5.2 Central Building

When predicting the energy consumption of the Central building, all the used models achieved identical error values per time horizon, Table 5.10. There were models that reached slightly higher accuracies, such as the SVM, XGB, and RF model for an hour, a day, and a week prediction horizon, respectively. In addition, from all the tested models, the highest CV(RMSE) error was, as in Civil building, observed for a week horizon, but now for the MLP model with a value of 13.96%.

Table 5.10 - Annual results for Central building forecast for an hour, a day, and a week horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units.

Models	1 hour			1 day			1 week		
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE
MLP	5.87	3.76	7.14	11.81	8.23	14.99	13.96	9.79	18.13
SVM	**4.96**	**2.94**	**5.74**	12.11	8.00	14.74	13.32	9.26	17.16
RF	5.31	3.14	6.17	11.13	7.36	**13.37**	**13.11**	**8.80**	**16.33**
XGB	5.09	3.12	6.05	**11.07**	**7.76**	14.04	13.63	9.44	17.33

Furthermore, from the monthly analysis performed with the prior selected models, Table 5.11, it is clear to state that when predicting one hour a-head consumption the SVM model achieved great predictions in every month, with CV(RMSE) values below 5.47%. On the other hand, for a day and a week horizon the selected models faced difficulties in predicting August when compared to other months. After all, for the three time horizons, two weeks of May were shown in Figure 5.7 as an example of the best case scenario, and two weeks of August were visualized as the worst case scenario in Figure 5.8.

Table 5.11 - Monthly results for Central building best models forecast by time horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units.

Months	1 hour - SVM			1 day - XGB			1 week - RF		
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE
January	4.80	2.93	5.71	11.99	7.76	14.52	13.66	8.26	16.35
February	5.17	3.07	6.23	**9.79**	6.89	12.77	12.22	7.66	15.18
March	5.37	2.98	6.39	10.72	7.87	14.60	12.57	8.86	17.14
April	4.16	2.58	4.85	10.66	8.08	13.99	11.46	8.88	15.55
May	**3.68**	**2.35**	**4.44**	9.80	6.87	**12.48**	**10.12**	7.74	**13.39**
June	4.83	3.16	5.88	12.51	8.68	15.85	15.96	9.80	18.66
July	4.75	3.10	5.94	10.98	8.04	14.86	11.91	9.66	17.38
August	4.75	2.91	5.26	13.87	8.60	15.41	17.64	10.21	19.28
September	5.44	3.12	6.73	10.06	**6.69**	13.59	13.17	**7.23**	16.73
October	5.25	3.04	5.91	10.07	6.99	12.68	12.36	7.89	14.82
November	5.47	3.03	6.05	10.67	7.84	13.62	10.76	8.64	14.52
December	5.33	2.99	5.53	11.48	8.75	13.98	14.13	10.59	16.88

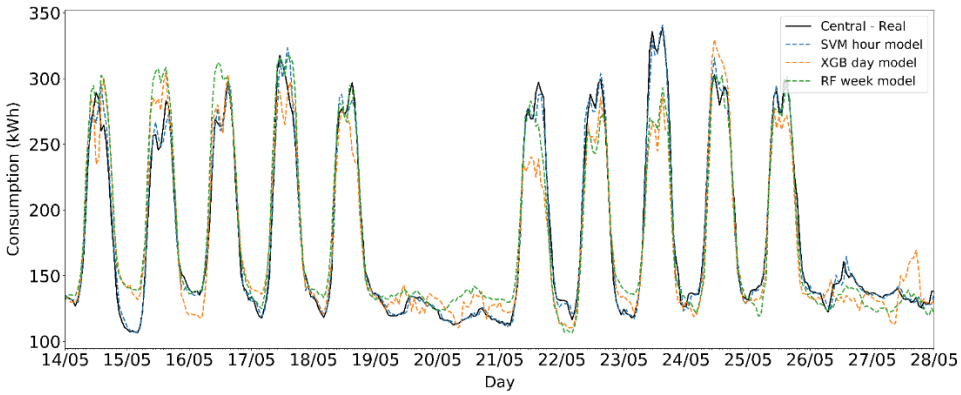


Figure 5.7 - Central building forecasting of two weeks in May for each time horizon best model

In the two weeks of May, Figure 5.7, as expected, the hour a-head forecasting had great accuracy, predicting almost the true energy consumption. At the same time, in a day and a week horizon prediction, the case was not the same, with a clearly decline of accuracy in some days of the week, as for example the twenty-third day of the month. In addition, it is also noticeable that for a day horizon forecasting the model tends to use the previous seen day to help in the prediction of the next one. An evident example of that occurs during the nighttime from fifteenth to sixteenth. It is also worth to mention, that this building base energy consumption is characterize by an unsteady behavior due to the 24 hours operating Data center, leading to less accurate predictions during nighttime when compared to the other building.

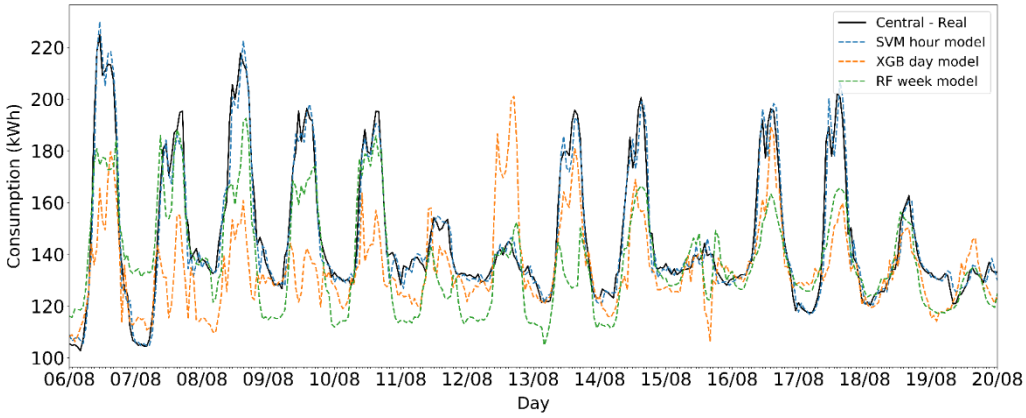


Figure 5.8 - Central building forecasting of the two weeks summer break for each time horizon best model

In the case of the two weeks summer break of August, Figure 5.8, the decrease of energy consumption that usually occurs in this time of the year was not as expected, probably due to some maintenance that was performed in the 24 hours operating Data center. As a consequence of that the day and the week horizon predictions were affected, registering smaller building base energy consumption than the true consumption. On contrary, the hour horizon predictions, was not influenced by that, fitting the true consumption almost perfectly.

Furthermore, in the day type results of Table 5.12, it is possible to conclude that in 93% of the days (workdays and weekends) the predicted consumptions did not exceed MAPE values of 3.15%, 9.31%, and 9.17% for an hour, a day, and a week horizons. In the remaining days the predictions of a day and a week horizon were not so good for the summer break, previous visualized in Figure 5.8, and for the 4% of holidays

Table 5.12 - Day type results for Central building best models forecast of each time horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units.

Day type	1 hour - SVM			1 day - XGB			1 week - RF			Day (%)
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	
Workdays	4.96	3.15	6.95	**10.02**	**6.62**	14.40	12.65	**8.14**	17.86	65
Weekends	**4.36**	**2.42**	**3.24**	12.92	9.31	**12.19**	**11.95**	9.17	**12.10**	28
Holidays	4.60	3.08	4.00	18.05	13.38	17.38	22.35	16.35	21.16	4
Summer break	4.61	2.69	4.40	17.16	10.32	17.65	15.88	9.70	16.06	3

5.5.3 North tower building

In North tower building annual results, Table 5.13, the model that represented each time horizon in the monthly and day type analysis was the SVM for every forecasting horizon. Moreover, in terms of the hour horizon prediction, it may be seen that SVM and MLP models outstood when compared with the other two DT based models. On the other hand, for a week horizon prediction, the SVM model showed higher capabilities against the other ones with improvements of above 2.5% in the CV(RMSE) metric.

Table 5.13 - Annual results for North tower building forecast for an hour, a day, and a week horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units.

Models	1 hour			1 day			1 week		
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE
MLP	10.19	6.16	**6.55**	18.87	13.84	13.58	22.60	14.34	16.70
SVM	**10.16**	**6.66**	6.74	**17.34**	**10.92**	**11.80**	**19.96**	**10.98**	**13.74**
RF	13.77	6.74	8.38	19.86	9.63	12.36	22.44	11.31	15.02
XGB	13.38	6.44	8.27	21.57	11.04	14.50	24.27	16.22	18.27

Moreover, with the prior selected models for each of the time horizons the monthly results were shown in Table 5.14. From that table, it is noticeable that the best month was not common to any of the time horizons tested, so since the models tend to increase the error with greater horizon predictions, the best month for the week horizon model, January, was selected to be visualized in Figure 5.9. Similar to the other buildings, the unpredictability of the two weeks summer break was also felt in the North tower building predictions, achieving the highest CV(RMSE) error for all the time horizons models. As a result of that, it was selected to be visualized as the worst case scenario of this building, Figure 5.10.

Table 5.14 - Monthly results for North tower building best models forecast by time horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units

Months	1 hour - SVM			1 day - SVM			1 week - SVM		
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE
January	9.32	5.93	5.31	12.91	7.65	7.00	**11.30**	**7.09**	**6.60**
February	8.77	5.93	**5.28**	11.98	**7.30**	**6.96**	12.72	7.64	7.86
March	7.91	**5.51**	5.40	**11.06**	7.79	7.90	15.28	7.82	9.95
April	14.08	7.39	8.27	22.86	13.15	15.19	24.62	13.82	17.44
May	9.41	6.22	7.04	19.50	10.59	13.67	21.91	10.45	16.58
June	9.31	6.66	6.83	18.21	13.36	14.95	22.83	15.76	18.36
July	8.51	6.19	6.30	14.13	11.14	12.30	14.08	10.36	12.40
August	18.20	8.23	8.23	27.97	14.62	14.63	31.63	15.82	16.61
September	**7.74**	6.13	6.42	13.87	9.86	11.58	16.99	8.88	14.31
October	8.15	6.64	7.18	13.59	9.61	11.40	19.38	9.26	15.37
November	8.38	6.62	6.88	15.28	10.47	12.69	17.40	9.37	13.58
December	11.72	8.27	7.47	21.40	14.84	12.61	22.41	14.98	14.85

In Figure 5.9, from the two weeks chosen for the best case scenario visualization, it is possible to check an almost similar weekly pattern from one week to another. That being said, since the greatest forecasting horizon used was one week, the models could easily predict this type of behaviour. In fact, that constant weekly pattern might explain the highest accuracy of the week horizon model over the day horizon model.

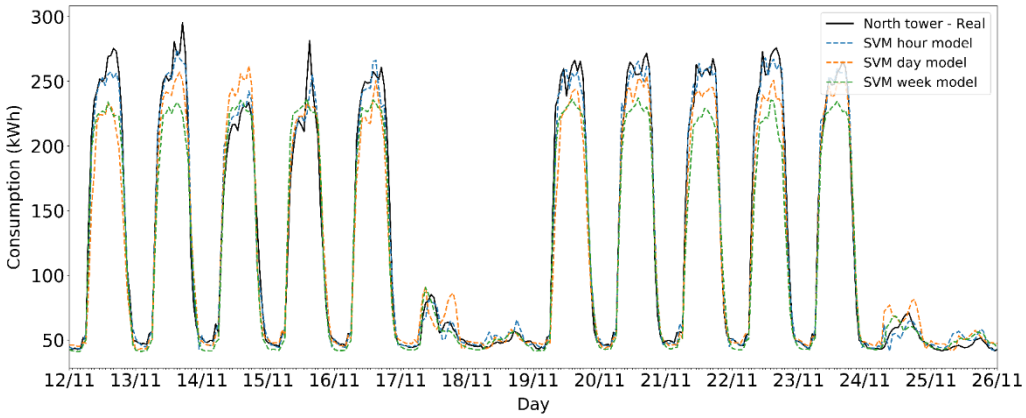


Figure 5.9 - North tower building forecasting of two weeks in January for each time horizon best model

From Figure 5.10, it is noticeable that in North tower building, the true energy consumption from the two weeks of summer break, did not fluctuate as much as in the Civil and Central building. For that reason, the day and week horizon models had achieved better predictions results in terms of the MAE metric than in the other two buildings, e.g. for the best week horizon model of Civil, Central, and North tower building, the MAE was 25.25%, 19.28%, and 16.61%, respectively. In addition, for the hour horizon model the predictions were not as good as in the other buildings.

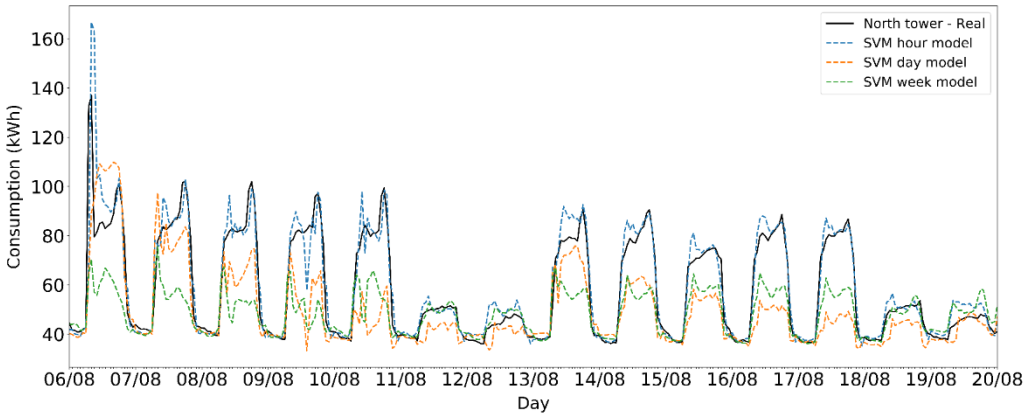


Figure 5.10 - North tower building forecasting of the two weeks summer break for each time horizon best model

From the day type results, Table 5.15, similar to the previous buildings, for all the horizons forecasted the models achieved greater accuracies when predicting the workdays and the weekends over the atypical days such as holidays and summer break days. Guaranteeing a MAPE below 16% for 93% of the days.

Table 5.15 - Day type results for North tower building best models forecast of each time horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units

Day type	1 hour - SVM			1 day - SVM			1 week - SVM			Day
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	(%)
Workdays	**8.86**	**5.45**	7.55	**14.30**	**7.84**	12.48	**17.67**	**10.26**	16.57	65
Weekends	16.15	9.12	4.98	30.94	15.93	**9.19**	25.17	10.48	**6.43**	28
Holidays	15.38	10.32	7.05	46.16	31.04	20.89	39.96	26.59	18.38	4
Summer break	13.96	6.91	**4.46**	31.60	16.35	12.12	37.95	16.84	13.31	3

5.5.4 South tower building

As last, in South tower building, the models that offered the most accurate annual results in the CV(RMSE) metric, Table 5.16, were the MLP for an hour horizon prediction and the XGB for a day and a week horizon prediction. It is also worth to refer the small deviations that occur in all the models results per time horizon prediction, which lead to the fact that none of the models selected presents the best results in the other two metrics (MAPE and MAE).

Table 5.16 - Annual results for South tower building forecast for an hour, a day, and a week horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units

Models	1 hour			1 day			1 week		
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE
MLP	**15.82**	7.02	13.33	21.01	11.92	20.89	23.81	12.59	23.33
SVM	16.11	7.24	13.29	21.06	**9.00**	**18.11**	23.68	10.62	21.42
RF	17.76	6.15	12.93	21.80	9.43	18.75	23.14	**10.21**	**20.82**
XGB	16.53	**6.08**	**12.59**	**20.81**	10.01	19.50	**22.31**	11.28	21.50

Furthermore, from the prior selected models the monthly errors were displayed in Table 5.17. According to that, it is noticeable that the months which achieved higher prediction values belong to the heating season, specifically, March, February, and November for an hour, a day, and a week horizon prediction. On the other hand, the cooling season included the lowest accurate months, being August the worst case scenario. That being said, the two weeks that were chosen to be visualized were from November and August.

Table 5.17 - Monthly results for South tower building best models forecast by time horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units

Months	1 hour - MLP			1 day - XGB			1 week - XGB		
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE
January	6.40	4.70	6.85	10.45	6.77	**9.90**	13.12	8.59	12.88
February	5.96	4.48	6.85	**10.03**	**6.51**	10.32	11.45	**7.44**	**11.81**
March	**5.74**	**4.03**	**6.29**	15.08	8.16	14.60	13.68	8.42	13.53
April	11.02	5.47	9.47	13.99	7.69	12.97	13.92	8.75	13.80
May	11.41	5.95	11.26	14.42	8.41	15.53	14.38	8.39	15.13
June	17.20	7.47	15.61	21.45	11.61	22.57	24.73	13.26	27.05
July	17.86	8.15	18.48	19.09	9.54	21.93	19.64	9.61	22.12
August	23.56	12.34	21.86	34.21	17.62	34.29	37.41	22.48	41.46
September	23.37	11.97	27.59	28.79	14.25	38.88	30.58	14.39	40.88
October	15.21	8.55	18.74	21.58	12.35	28.12	22.76	14.46	31.98
November	8.90	5.74	9.92	12.43	8.82	13.89	**10.87**	8.05	12.64
December	7.67	5.47	7.39	13.99	8.35	11.42	18.78	11.46	15.33

In one of the best month of South tower building prediction, Figure 5.11, it this possible to encounter the same kind of constant weekly pattern found in North tower building best case scenario example, in Figure 5.9. As a consequence of that pattern the week horizon model was more precise than the day horizon model, which might be explained by the importance that each model gives to the features that it is provided with. In this particular case, since the week horizon model just has the consumption from one week before the prediction, its decision of what might be the next forecasting value is more affected by it, in comparison with the day horizon model, which has the day and the week before the actual prediction.

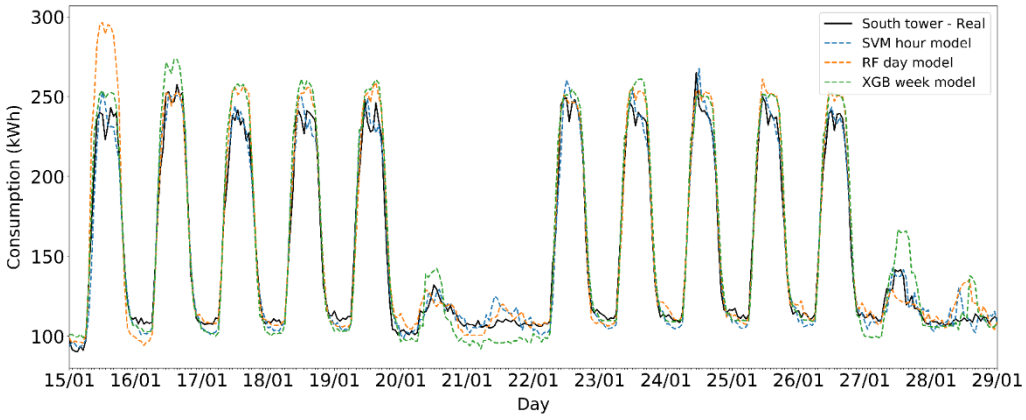


Figure 5.11 - South tower forecasting of two weeks of November for each time horizon best model

In the two weeks of summer break, visualized in Figure 5.12, the predictions results are somehow similar to what was observed in the North tower building, Figure 5.10. Although the true consumption was quite different, with sudden peaks of energy during the weekends and almost three times more the mean energy consumption during the daytime of the days of the week.

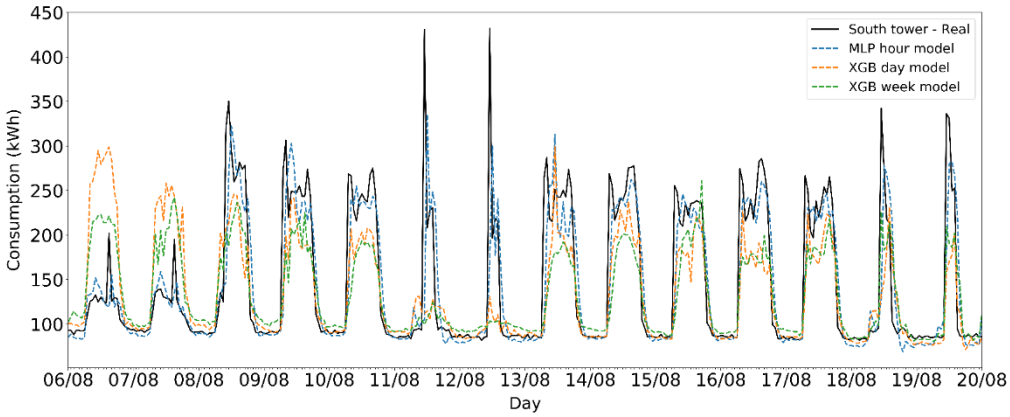


Figure 5.12 - South tower building forecasting of the two weeks summer break for each time horizon best model

Once again, from Table 5.18, the two types of day that yield better results in most of the horizon models, was the workday followed by the weekends, which together account with 93% of all the days forecasted and achieved MAPE values inferior to 12.7%.

Table 5.18 - Day type results for South tower building best models forecast of each time horizon, where CV(RMSE) and MAPE are presented in percentage and MAE is presented in kWh units.

Day type	1 hour - MLP			1 day - XGB			1 week - XGB			Day
	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	CV(RMSE)	MAPE	MAE	(%)
Workdays	**13.00**	**6.33**	14.06	**17.75**	**8.19**	19.85	**19.70**	**9.71**	22.57	65
Weekends	26.08	7.96	11.21	29.17	11.77	**16.19**	29.40	12.66	**16.81**	28
Holidays	29.37	13.31	22.87	40.10	19.47	36.03	38.46	20.40	37.04	4
Summer break	14.36	8.63	**10.61**	29.76	17.73	21.68	31.09	17.90	22.11	3

5.5.5 Complementary Visualization of Atypical Weeks

In addition to what was former analysed, two atypical weeks were visualized for each of the buildings in order to understand the models' ability to adapt. One of the weeks corresponds to the carnival period and the other one to the last week of the year, where holidays (Christmas and New Year) took place. The choice of these two weeks was supported by the degree of difficulty that is added from one week to another. The first being the easiest to predict with just one holiday, Figure 5.13, and the second being the hardest with three different holidays, Figure 5.14.

For the carnival week starting at twelfth of February, Figure 5.13, it may be seen that the behave of each forecasting horizon model in most of the buildings was satisfactory. Which means, that the $s\_workday$ feature for the holiday on the thirteenth day of the month worked as expected, decreasing buildings energy consumption in a presence of an holiday, in the day and week horizon models that did not had any reference about the previous hour consumption.

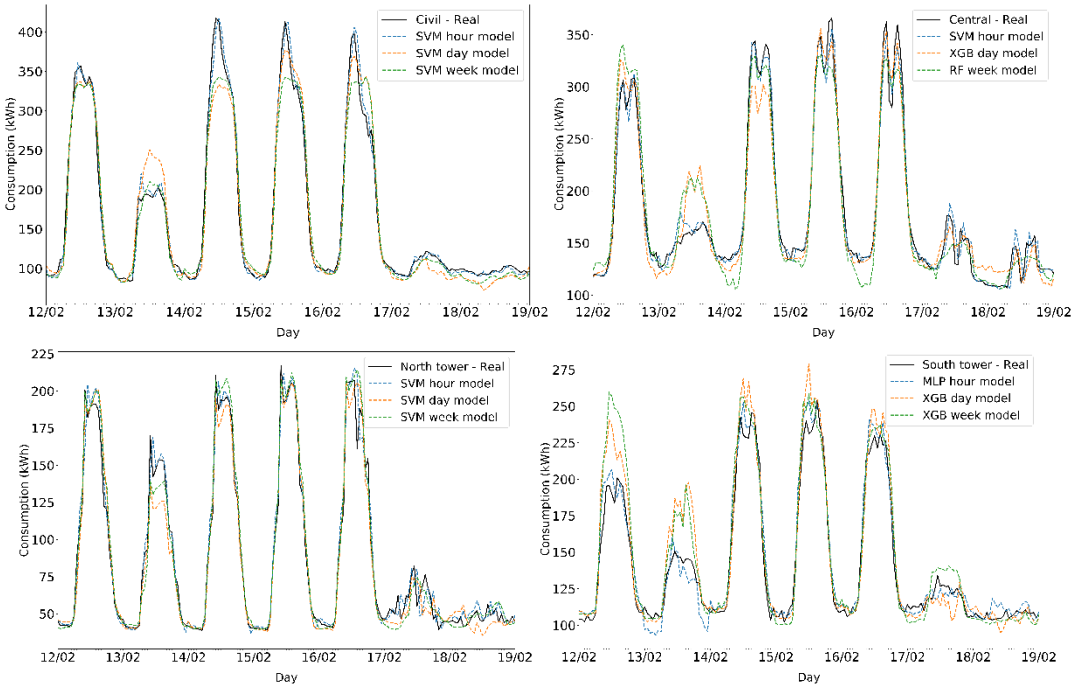


Figure 5.13 - Carnival week predictions for each building and time horizon best model

On the other hand, in the last week of the year, Figure 5.14, the presence of more holidays creates an extra stochastic behaviour of each buildings occupants, leading to poor predictions in most of the buildings for the day and week horizon models. Among all the building, South tower building was the worst one achieving absurd consumptions during the Christmas day.

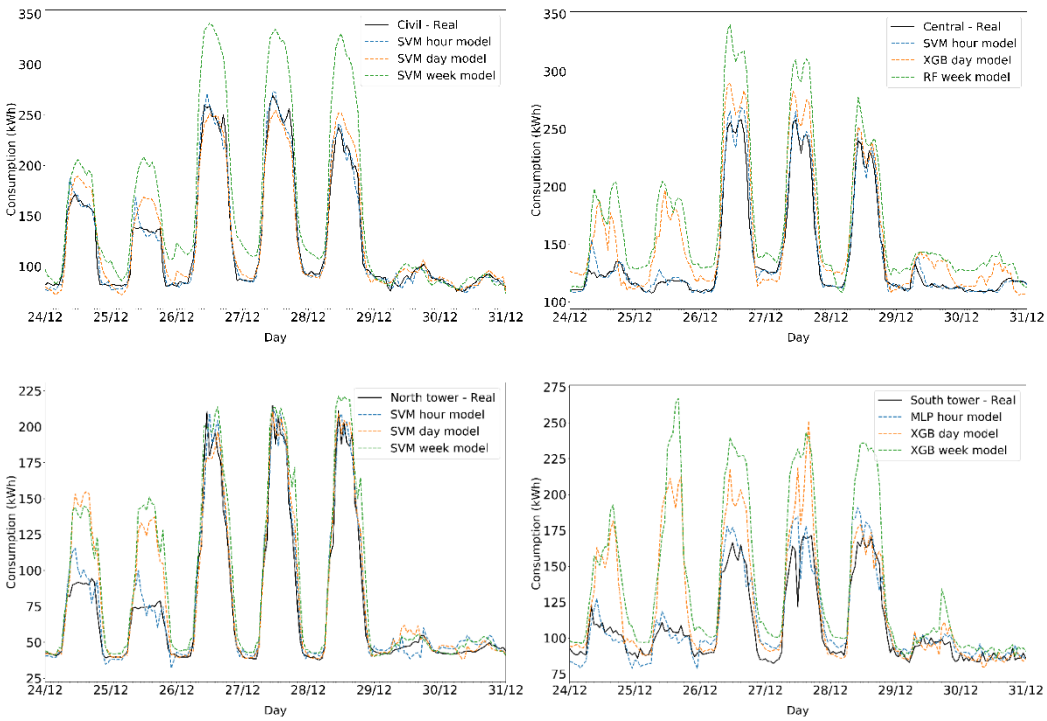


Figure 5.14 - Last week of the year predictions for each building and time horizon best model

Chapter 6

Conclusions

In this work, four machine learning models (MLP, SVM, RF, and XGB) were compared in three different forecasting horizons (an hour, a day, and a week) for four main buildings (Civil, Central, North tower, and South tower) located at Instituto Superior Técnico, Lisbon, giving a total of 48 models developed. To conduct this study, two types of hourly collected data (WCD and ECD) for three years (2014, 2017, and 2018) were used. In order to create the right conditions for each forecasting model development, the available data was first treated and then analysed.

In the first stage, a data imputation study was performed, concluding that from the two multiple imputation algorithms used (MICE and MF), the one that achieved greater results was the MF algorithm, which was successfully applied in each building consumption data incompleteness (ECD). However, in the presence of simultaneous wide gaps of missing values, which occurred in WCD, this algorithm was not able to supply the tendency and seasonality intrinsic to the type of variable that was being filled. For that reason, a single imputation method created with that propose (HMM) was employed, guaranteeing with that the necessary dynamic of the imputed variable.

In the second stage of this work, each building energy consumption was analysed and several features were created in an attempt to better define the buildings' behaviour. With all the features generated, a feature selection analysis via the RFE method took place to understand the importance of each feature. From that analysis, it was concluded that three type of features were indispensable in every building and forecasting horizon data set, specifically, the day type, the lagged features (one hour, one day, and one week), and the cluster average consumption of the larger group of days of each building from the year of 2017. In addition, it was also noticeable that the used of this method was too restrict in the feature selection process, which might be explained by the use of the clusters average consumption features that were generated from one of the years where this method was applied (training set), influencing with that the importance of each feature and the consequent selection. To overcome that situation, a new selection supported by the RFE method selection was employed.

Furthermore, after the hyperparameter optimization and the forecast results obtained, it was concluded that, even though all the models developed did not show larger error variations when predicting the same building consumption and forecasting horizon, SVM model outstood, achieving the most accurate results in the majority of the predictions performed. One of the reasons which might explain this situation is supported by the small number of hyperparameters that this model has, leading it to reach its full potential faster and more easily than any other model used in this work. Moreover, the second most accurate model was XGB, performing the best prediction in three of the cases, followed

by the MLP and RF that achieved the highest accuracy value in one case each. Considering the best models' selection, in most of the buildings and forecasting horizons it was found that August was the hardest month to predict, due to the intrinsic unpredictability of two summer break weeks of all Alameda's campus facilities. On the other hand, in 93% of the predicted days, which accounts with working days (65%) and weekends (28%) the buildings achieved a MAPE error of 10.95%, 9.17%, 10.48%, and 12.66% for Civil, Central, North tower, and South tower buildings in a week horizon prediction. In addition to that, an increasing annual error tendency was noticeable, when the models attempt to predict in greater horizons.

Finally, using the CV(RMSE) metric to compare the different buildings predictions, it was found that in every forecasting horizon studied, Central building was the easiest one to predict and South tower building the hardest one. This might be related by the small variations of consumption found in the first building, against the sudden peaks and dips encounter in the latter.

6.1 Future Work

In this work, it was concluded that the stochastic behaviour of each building's inhabitants is one of the major reasons for the variations of the energy consumption patterns. With that in mind, two solutions can be adopted:

- The simplest one, is to perform a different day type feature for each of the buildings, since it was shown that different buildings have different consumption patterns in the same type of day;
- The second solution, it only can be applied in buildings with Wi-Fi systems, with that being said, it basically uses each people network log in to count the online number of occupants in each of the buildings.

As it is known, machine learning models are widely dependent on data and one of the limitations of this study was the quantity of data and the years that were available. The fact that just two years were available, and they were so apart from each other (2014 and 2017), it influenced the learning process of the algorithms to predict the year of 2018. To address that situation, the average cluster features were created based only on 2017 year consumptions, although further improvements may be done such as:

- Using an attribute called warm start, that is present in a vast majority of the machine learning models used, like MLP and RF. This attribute allows the division of the training process. As future work, the learning process could be divided into two. The first stage will train in the 2014 and 2017 years and save each model learned parameters to the next stage. The second stage, with the parameters already "warmed", will use only 2017 (the year closest to the one to predict) to learn better the patterns that are more similar to the year to predict.

Another future work that may be explored is supported by the fact that each model achieved similar error measurements. With that, an ensemble model with an evolutionary algorithm to define each model percentage in the final output may be employed, which, normally, leads to a generalized and enhanced prediction.

Lastly, it has been shown in several recent studies the tendency of using deep learning models to predict sequential data, such as time series. The use of those models such as long-short term memory (LSTM) could be beneficial, due to its intrinsic long term dependencies in recurrent architectures.

References

[1] E. Roser, Max, Ortiz-Ospina, "World Population Growth," *Published online at OurWorldInData.org*, 2017. [Online]. Available: https://ourworldindata.org/world-population-growth#population-growth.

[2] M. A. Z. and A. F. S. María V. Moreno, "User-centric smart buildings for energy sustainable smart cities," *Wiley Online Library*, 2013. .

[3] "Cities." [Online]. Available: https://ec.europa.eu/clima/policies/international/paris_protocol/cities_en. [Accessed: 28-Nov-2018].

[4] M. Leahy, J. L. Barden, B. T. Murphy, N. Slater-thompson, and D. Peterson, "International Energy Outlook 2013."

[5] European Commission, "Buildings - European Commission," 2018. [Online]. Available: https://ec.europa.eu/energy/en/topics/energy-efficiency/buildings. [Accessed: 30-Nov-2018].

[6] Eurostat, "2014 energy consumption by sector in the EU," 2014. [Online]. Available: https://epthinktank.eu/2016/07/08/energy-efficiency-in-buildings/energy-consuption-by-sector/. [Accessed: 06-Dec-2018].

[7] E. Commission, "European Commission, Action Plan for Energy Efficiency: Realizing the Potential, Communication from the Commission," *Eur. Comm. Action Plan Energy Effic. Realiz. Potential, Commun. from Comm.*, 2006.

[8] M. Paridah, A. Moradbak, A. . Mohamed, F. abdulwahab taiwo Owolabi, M. Asniza, and S. H. . Abdul Khalid, "We are IntechOpen , the world ' s leading publisher of Open Access books Built by scientists , for scientists TOP 1 %," *Intech*, vol. i, no. tourism, p. 13, 2016.

[9] 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.

[10] S. Pan *et al.*, "A review of data-driven approaches for prediction and classification of building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 1027–1047, Feb. 2017.

[11] M. S. Al-Homoud, "Computer-aided building energy analysis techniques," *Build. Environ.*, vol. 36, no. 4, pp. 421–433, May 2001.

[12] C. S. Barnaby and J. D. Spitler, "Development of the residential load factor method for heating and cooling load calculations," in *ASHRAE Transactions*, 2005, vol. 111 PART 1, pp. 291–307.

[13] J. Cavalheiro and P. Carreira, "A multidimensional data model design for building energy management," *Adv. Eng. Informatics*, vol. 30, no. 4, pp. 619–632, Oct. 2016.

[14] Z. Li, Y. Han, and P. Xu, "Methods for benchmarking building energy consumption against its past or intended performance: An overview," *Applied Energy*, vol. 124. Elsevier, pp. 325–334,

01-Jul-2014.

[15] S. Paudel, P. H. Nguyen, W. L. Kling, M. Elmitri, B. Lacarrière, and O. Le Corre, “Support Vector Machine in Prediction of Building Energy Demand Using Pseudo Dynamic Approach,” Jul. 2015.

[16] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, “Big Data Analytics for Dynamic Energy Management in Smart Grids,” *Big Data Res.*, vol. 2, no. 3, pp. 94–101, Sep. 2015.

[17] H. X. Zhao and F. Magoulès, “A review on the prediction of building energy consumption,” *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6. pp. 3586–3592, 2012.

[18] “FDelca/energy_consumption_forecasting.” [Online]. Available: https://github.com/FDelca/energy_consumption_forecasting. [Accessed: 31-Oct-2019].

[19] D. Kolokotsa, “The role of smart grids in the building sector,” *Energy Build.*, vol. 116, pp. 703–708, 2016.

[20] M. Manic, D. Wijayasekara, K. Amarasinghe, and J. J. Rodriguez-Andina, “Building Energy Management Systems: The Age of Intelligent and Adaptive Buildings,” *IEEE Ind. Electron. Mag.*, vol. 10, no. 1, pp. 25–39, 2016.

[21] P. Waide, J. Ure, N. Karagianni, G. Smith, and B. Bordass, “The scope for energy and CO2 savings in the EU through the use of building automation technology,” 2013.

[22] F. Rabhi *et al.*, “An overview of the commercial cloud monitoring tools: research dimensions, design issues, and state-of-the-art,” *Computing*, vol. 97, no. 4, pp. 357–377, Apr. 2014.

[23] J. Sides, “The Victory Lab: The Secret Science of Winning Campaigns,” *Public Opin. Q.*, vol. 78, no. S1, pp. 363–364, Jan. 2014.

[24] W. J. Kuo, R. F. Chang, D. R. Chen, and C. C. Lee, “Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images,” *Breast Cancer Res. Treat.*, vol. 66, no. 1, pp. 51–57, Mar. 2001.

[25] M. Qamar and A. Khosravi, “A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings,” *Renew. Sustain. Energy Rev.*, vol. 50, pp. 1352–1372, 2015.

[26] K. Amasyali and N. M. El-Gohary, “A review of data-driven building energy consumption prediction studies,” *Renewable and Sustainable Energy Reviews*, vol. 81. 2018.

[27] G. R. Newsham and B. J. Birt, “Building-level occupancy data to improve ARIMA-based electricity use forecasts,” in *BuildSys*, 2010, p. 13.

[28] S. R. Twanabasu and B. A. Bremdal, “Load forecasting in a smart grid oriented building,” in *22nd International Conference and Exhibition on Electricity Distribution (CIRED 2013)*, 2013, pp. 0907–0907.

[29] N. Mohamed, M. H. Ahmad, Suhartono, and Z. Ismail, “Improving short term load forecasting using double seasonal arima model,” *World Appl. Sci. J.*, vol. 15, no. 2, pp. 223–231, 2011.

[30] J. Zhuang, Y. Chen, X. Shi, and D. Wei, “Building Cooling Load Prediction Based on Time Series Method and Neural Networks,” *Int. J. Grid Distrib. Comput.*, vol. 8, no. 4, pp. 105–114, 2015.

[31] Y. K. Penya, C. E. Borges, D. Agote, and I. Fernández, “Short-term load forecasting in air-conditioned non-residential Buildings,” in *Proceedings - ISIE 2011: 2011 IEEE International Symposium on Industrial Electronics*, 2011, pp. 1359–1364.

[32] J. Colomer, L. Burgas, J. Massana, C. Pous, and J. Melendez, “Short-term load forecasting in a non-residential building contrasting models and attributes,” *Energy Build.*, vol. 92, pp. 322–330, Apr. 2015.

[33] D. Zhao, M. Zhong, X. Zhang, and X. Su, “Energy consumption predicting model of VRV (Variable refrigerant volume) system in office buildings based on data mining,” *Energy*, vol. 102, pp. 660–668, May 2016.

[34] J. Miguel, “Desenvolvimento de modelos de estimação de consumos em edifícios de serviços para implementação de alarmística de gestão de energia,” 2016.

[35] M. Borovcnik, H.-J. Bentz, and R. Kapadia, “A Probabilistic Perspective,” in *Chance Encounters: Probability in Education*, 1991, pp. 27–71.

[36] Z. Zhao, S. Xu, B. H. Kang, M. M. J. Kabir, Y. Liu, and R. Wasinger, “Investigation and improvement of multi-layer perception neural networks for credit scoring,” *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3508–3516, 2015.

[37] P. A. González and J. M. Zamarreño, “Prediction of hourly energy consumption in buildings based on a feedback artificial neural network,” *Energy Build.*, vol. 37, no. 6, pp. 595–601, Jun. 2005.

[38] S. Eric, “Predicion and control using feedback neural networks and partial models,” 1996.

[39] S. Karatasou, M. Santamouris, and V. Geros, “Modeling and predicting building’s energy use with artificial neural networks: Methods and results,” *Energy Build.*, vol. 38, no. 8, pp. 949–958, Aug. 2006.

[40] I. Rivals and L. Personnaz, “Neural-network construction and selection in nonlinear modeling,” *IEEE Trans. Neural Networks*, vol. 14, no. 4, pp. 804–819, Jul. 2003.

[41] R. Yokoyama, T. Wakui, and R. Satake, “Prediction of energy demands using neural network with model identification by global optimization,” *Energy Convers. Manag.*, vol. 50, no. 2, pp. 319–327, 2009.

[42] R. Yokoyama and K. Ito, “Capability of Global Search and Improvement in Modal Trimming Method for Global Optimization,” *JSME Int. J. Ser. C*, vol. 48, no. 4, pp. 730–737, 2006.

[43] G. Escrivá-Escrivá, C. Álvarez-Bel, C. Roldán-Blay, and M. Alcázar-Ortega, “New artificial neural network prediction method for electrical consumption forecasting based on building end-uses,” *Energy Build.*, vol. 43, no. 11, pp. 3112–3119, Nov. 2011.

[44] C. Roldán-Blay, G. Escrivá-Escrivá, C. Álvarez-Bel, C. Roldán-Porta, and J. Rodríguez-García,

"Upgrade of an artificial neural network prediction method for electrical consumption forecasting using an hourly temperature curve model," *Energy Build.*, vol. 60, pp. 38–46, May 2013.

[45] R. Mena, F. Rodríguez, M. Castilla, and M. R. Arahal, "A prediction model based on neural networks for the energy consumption of a bioclimatic building," *Energy Build.*, vol. 82, pp. 142–155, Oct. 2014.

[46] R. Platon, V. R. Dehkordi, and J. Martel, "Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis," *Energy Build.*, vol. 92, pp. 10–18, Apr. 2015.

[47] K. Li, C. Hu, G. Liu, and W. Xue, "Building's electricity consumption prediction using optimized artificial neural networks and principal component analysis," *Energy Build.*, vol. 108, pp. 106–113, Dec. 2015.

[48] 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, Jan. 2016.

[49] K. Li, X. Xie, W. Xue, X. Dai, X. Chen, and X. Yang, "A hybrid teaching-learning artificial neural network for building electrical energy consumption prediction," *Energy Build.*, vol. 174, pp. 323–334, 2018.

[50] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[51] F. Zhang, C. Deb, S. E. Lee, J. Yang, and K. W. Shah, "Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique," *Energy Build.*, vol. 126, pp. 94–103, Aug. 2016.

[52] Q. Li, Q. Meng, J. Cai, H. Yoshino, and A. Mochida, "Applying support vector machine to predict hourly cooling load in the building," *Appl. Energy*, vol. 86, no. 10, pp. 2249–2256, Oct. 2009.

[53] X. Li, J. H. Lǔ, L. Ding, G. Xu, and J. Li, "Building cooling load forecasting model based on LS-SVM," *Proc. - 2009 Asia-Pacific Conf. Inf. Process. APCIP 2009*, vol. 1, pp. 55–58, 2009.

[54] X. Li, L. Ding, and L. Li, "A novel building cooling load prediction based on SVR and SAPSO," *3CA 2010 - 2010 Int. Symp. Comput. Commun. Control Autom.*, vol. 1, no. 1, pp. 528–532, 2010.

[55] Y. Fu, Z. Li, H. Zhang, and P. Xu, "Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-metering Devices," *Procedia Eng.*, vol. 121, pp. 1016–1022, 2015.

[56] L. K. Hansen and P. Salamon, "Neural Network Ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, 1990.

[57] D. Opitz and R. Maclin, "Popular Ensemble Methods: An Empirical Study," *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, Jul. 1999.

[58] Z. Wang and R. S. Srinivasan, “A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models,” *Renewable and Sustainable Energy Reviews*, vol. 75, no. September 2015. Elsevier Ltd, pp. 796–808, 2017.

[59] C. Fan, F. Xiao, and S. Wang, “Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques,” *Appl. Energy*, vol. 127, pp. 1–10, 2014.

[60] 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.

[61] Z. Wang, Y. Wang, R. Zeng, R. S. Srinivasan, and S. Ahrentzen, “Random Forest based hourly building energy prediction,” *Energy Build.*, vol. 171, pp. 11–25, 2018.

[62] C. Robinson *et al.*, “Machine learning approaches for estimating commercial building energy consumption,” *Appl. Energy*, vol. 208, no. May, pp. 889–904, 2017.

[63] "Basic Ensemble Learning (Random Forest, AdaBoost, Gradient Boosting)- Step by Step Explained", Medium, 2019. [Online]. Available: https://towardsdatascience.com/basic-ensemble-learning-random-forest-adaboost-gradient-boosting-step-by-step-explained-95d49d1e2725. [Accessed: 11- Nov- 2019]

[64] S. van Buuren and K. Oudshoorn, “Flexible multivariate imputation by MICE,” pp. 1–20, 1999.

[65] D. J. Stekhoven and P. Bühlmann, “Missforest-Non-parametric missing value imputation for mixed-type data,” *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[66] Python Software Foundation, “Welcome to Python.org,” *2001*, 2017. [Online]. Available: https://www.python.org/. [Accessed: 19-Oct-2019].

[67] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19. pp. 2507–2517, 01-Oct-2007.

[68] C. Jeffery, “‘To See the World in a Grain of Sand’: Wolfgang Laib and the Aesthetics of Interpenetrability,” *Relig. Arts*, vol. 17, no. 1–2, pp. 57–73, 2013.

[69] “skopt API documentation,” *scikit-optimize.github.io*. [Online]. Available: https://scikit-optimize.github.io/. [Accessed: 14-Oct-2019].

[70] “scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation.” [Online]. Available: https://scikit-learn.org/stable/index.html. [Accessed: 30-Oct-2019].

[71] Xgb. Developers, “XGBoost Documentation,” *XGBoost*, 2016. [Online]. Available: https://xgboost.readthedocs.io/en/latest/. [Accessed: 30-Oct-2019].

[72] "missingpy", *PyPI*, 2019. [Online]. Available: https://pypi.org/project/missingpy/. [Accessed: 31- Oct- 2019]

[73] M. Ana, "Plataforma de gestão de energia para o Campus do IST: Modelação e representação

dos consumos de energia", Instituto Superior Técnico, 2019.