1 Introduction

1.1 Motivation

With the increasing concern about the climate change and carbon emissions, it is of primordial importance to identify the sectors and activities that contribute the most to this effect, so corrective and preventive measures can be implemented.

The first activities that come to mind when thinking about this issue are the industry and the transportation, which is totally normal. Surely because we are accustomed to hearing and reading news about this topic in the media. Local and state governments pass laws that aim to mitigate the harmful impact of these activities, like restricting traffic to certain cars on especially bad days and penalizing or even closing down polluting factories. News of all this quickly reach the citizens ears.

But, the activity that almost no one thinks about when asked about this topic is actually the most energy consuming one, making it almost the most polluting as well. The answer is quite shocking, or at least it was for the authors of this report. It is buildings.

According to the European Commission [1], buildings, both residential and non-residential, are responsible for approximately 40% of energy consumption and 36% of CO_2 emissions in the EU. This makes them the single largest energy consumer, followed by the industry, and the second emitter of greenhouse effect gases (GHG), behind the transportation.

This obviously constitutes a big issue as it needs to be approached and solved without putting in harm's way the people's quality of life. This is, we must still be able to heat the building in winter, cool it in summer and, in general, perform any activity necessary to lead a comfortable life. Also, to add to this, reports suggest that building energy consumption has steadily increased over the past decade, being heating, ventilation and air conditioning (HVAC) the factors that account for most of the energy use.

The European Union is aware of this, that is why, inspired by the Kyoto Protocol, in December of 2002 the directive on the energy performance of buildings (EPBD, Energy Performance of Buildings Directive) was stablished, aiming to promote the improvement of the energy performance and energy conservation properties of buildings within the community.

As mentioned above, the EPDB promotes policies that will help mitigate the impact of this issue, by focusing both on existing and newly built buildings.

Regarding the first group, nowadays, about 35% of the EU's buildings are over 50 years old and almost 75% are energy inefficient, that is why it is necessary to renovate them. By doing this, the energy is used in a more efficient way, reducing their environmental footprint and, additionally, helping create economic and social benefits, contributing to improving the health, comfort and wellbeing of the residents, and making homes more affordable. The goal is that all European countries implement a national roadmap to ensure the complete decarbonization of the building stock by 2050.

And, for the second group, the aim is that all new constructions should be nearly zero-energy buildings (NZEB) as of 31 December 2020. These building incorporate new technologies and improvements that make them have a very high energy performance, and the low amount of energy that they do require comes mostly from renewable sources, making their environmental footprint almost nonexistent.

1.2 Focus of this report

When approaching the task of designing an efficient building, there are a multitude of factors that must be considered and, also, the problem can be tackled from multiple directions, each one corresponding to a particular field of science.

To give some examples, a building could be made more efficient by making them smarter, meaning adding automation and control systems that adapt to the needs of the consumer and interact with the grid. They could be built using materials with better thermal and insulating properties. Solar panels or other renewable energy devices can be installed so to produce clean electricity.

Also, it's important to mention that factors such as the climate of the region where the building is being constructed, the intended use (residential/industrial) and the occupancy must also be taken into account when implementing solutions as the ones just mentioned.

As it can be deduced from all this, there are almost an endless number of ways the design of a building can be modified to make it more efficient, but on this report the focus falls on two factors that haven't been mentioned yet. The shape or form of the building and its orientation.

On the following sections we tried to answer the question of how these factors impact the efficiency of a building from a statistical point of view. As previously mentioned, this question is quite important, particularly in the European Union, and must be solved with as much accuracy as possible.

1.3 Industry 4.0

Lastly, before we start with the statistical analysis, we want to frame this problem on the context of Industry 4.0.

Even though the construction of efficient buildings is not one of the main premises of Industry 4.0, companies are starting to understand that they need to take part in the quest to reduce polluting emissions and reduce energy consumption. That is why most of them are trying to make their installations more environmentally friendly, be it their offices or factories.

With this, they not only save energy and reduce emissions, but they also save water and produce less waste. Furthermore, another consequence of all this is that the view that the general public and the world has of them also improves quite significantly.

2 Statistical Analysis

2.1 The data

As stated above, the main objective of this work is to establish how the shape and orientation of a building influences its energetic efficiency or, to be more precise, how it influences two magnitudes, the heating load (CL) and cooling load (HL) [2].

The heating load refers to the amount of heat energy that would need to be added to a space to maintain the temperature in acceptable range. And, likewise, the cooling load is the amount of heat energy that would need to be removed from a space to maintain the temperature in an acceptable range. So, as it can be interpreted, lower "thermal loads" indicate that, relatively, the building will require less heating and cooling to maintain comfortable conditions.

Both these magnitudes take into account the building's construction an insulation, including floors, walls, ceiling and roof, and the building glazing and skylights, based on size, performance, shading and overshadowing. These last points are the ones subject to change with the shape and orientation of the building.

The HL and CL are also really important in order to determine the appropriate HVAC equipment the building is going to need.

To carry out the analysis of these two magnitudes we used simulated data, obtained from a building energy simulation tool called Autodesk Ecotet Analysis [3].

With the help of this program, 728 possible buildings were created. First, by using an elementary cube of 3.5 m each side and arranging 18 of them, 12 different shapes were created, each of them with different dimensions and surface areas, but with the same volume (771.75 m^3). Lastly, by changing the amount of glazing area[1] and its orientation, the rest of the combinations were obtained. The construction materials are also the same for all buildings.

Then, by setting the desired indoor conditions, this is, humidity (60%) and temperature (19-24°C), the simulation program obtains the estimated heating load (*Y1*) and cooling load (*Y2*) for each building as a function of eight numeric variables that act as predictors or inputs. They are:

- *X1*: Relative compactness (12)*
- *X2*: Surface area (12)
- *X3*: Wall area (7)
- *X4*: Roof area (4)
- *X5*: Overall height (2)
- *X6*: Orientation (4)
- *X7*: Glazing area (4)
- *X8*: Glazing area distribution (6)

* Number of possible values the corresponding variable can take.

In the next sections of this report, the statistical analysis of the data was carried out. Different linear regression and classification techniques were used, with the final objective of ascertaining the impact that each of the variables has on the thermal loads and obtaining models that allow to predict these magnitudes for different shapes and orientations.

But, before that, a brief preliminary analysis is conducted. By doing this it is possible to determine which are the best techniques to use and expected results in relation with the statistical properties of the data.

[1] The glazing area refers to the percentage of surface area of the building that is made of glass.

2.2 Preliminary Analysis

The first step to take when analyzing a dataset is to conduct an unbiassed exploration of the statistical properties of both the target and predictor variables (histograms and density plots) and their relationship (correlation and scatterplots).

The histograms for the heating load (*Y1*) and cooling load (*Y2*) variables can be seen on *Figure 1*. They show that both variables span a continuous range of values and, for different ranges, we can obtain the number of appearances in the data.

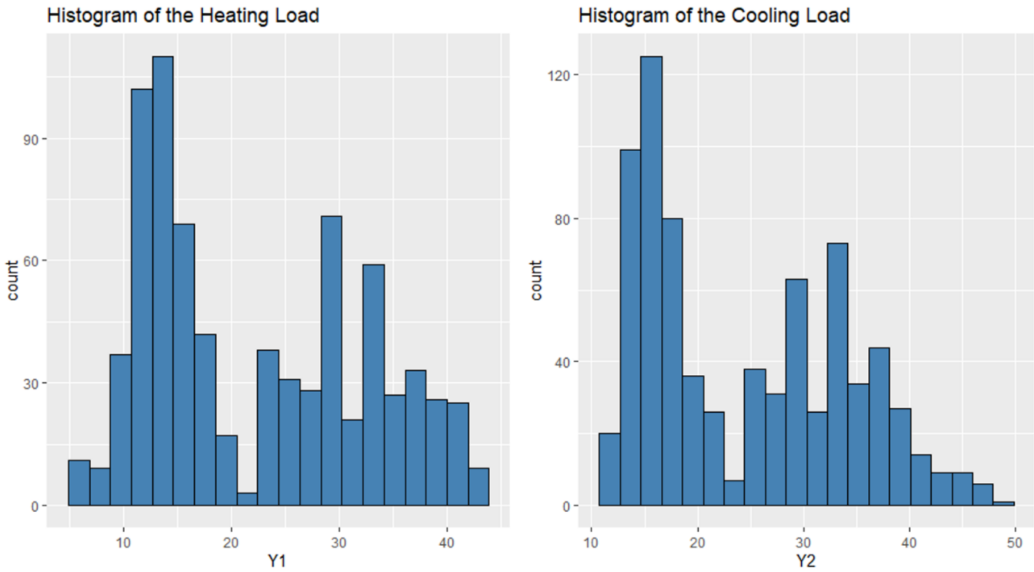


Figure 1 - Histograms of the HL and CL

Figure 2 contains the density plots for all variables. They show the distribution followed by said variables and, as it is possible to see, neither of them follows a normal or Gaussian distribution.

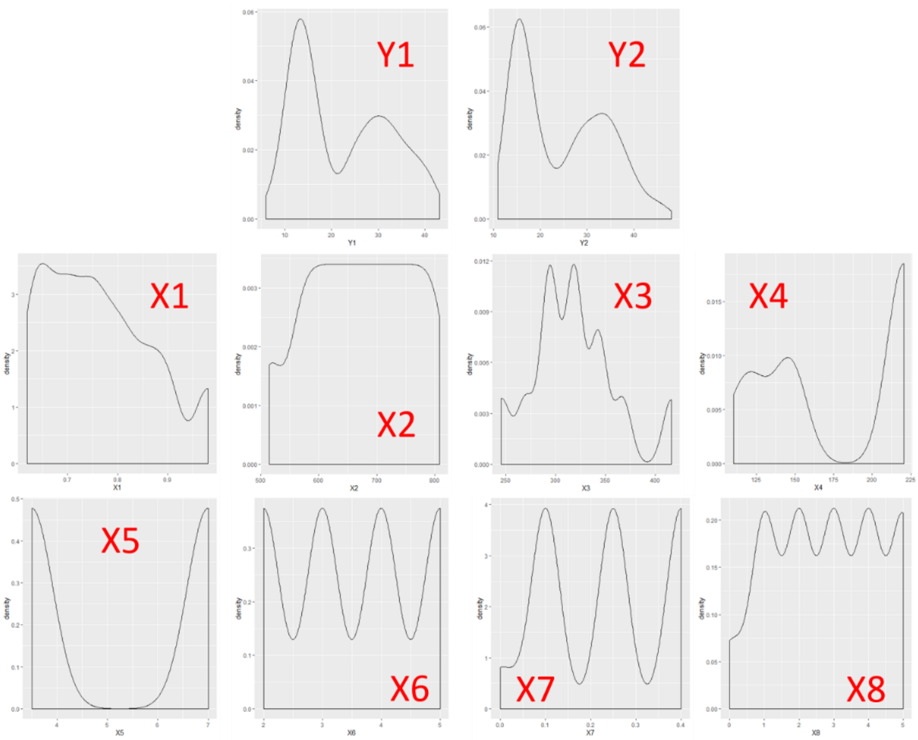


Figure 2 - Density plots

Figure 3 display the scatterplots for each of the predictor or input variables with the heating load (the scatter plots for the cooling load have been omitted as they are quite similar). These scatterplots show that the relationship between the input and output variables is far from trivial, as linear patterns are difficult to discern in some cases.

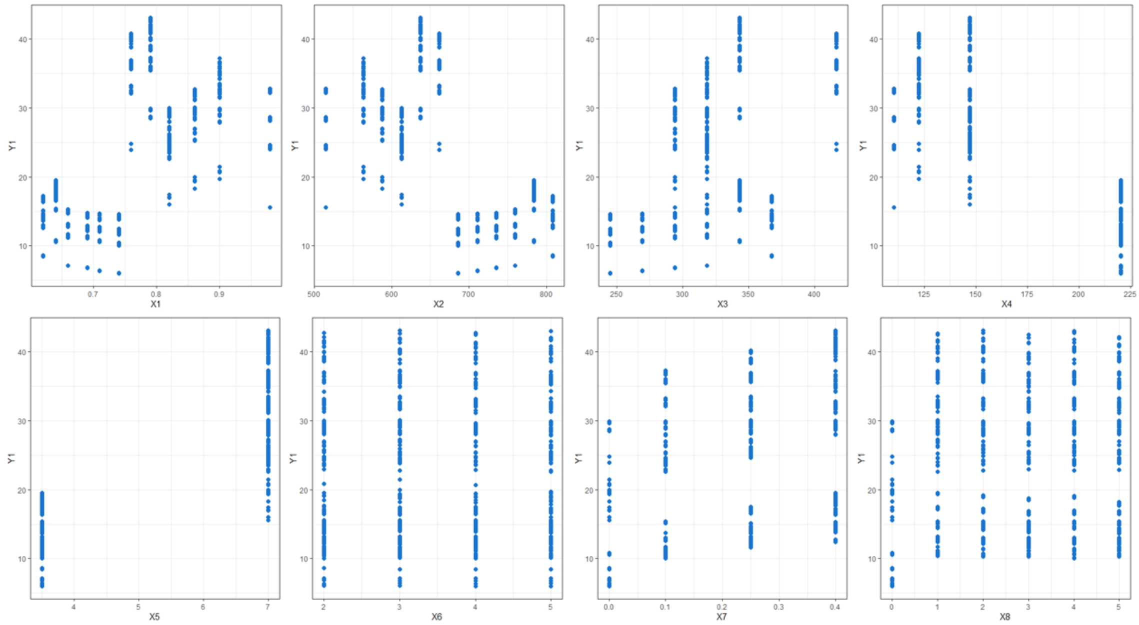


Figure 3 - Scatter plots between the heating load and predictor variables

Lastly, *Figure 4* shows the correlation between all variables. In the case of the output variables, *Y1* and *Y2*, the input variables that are more strongly correlated (negative or positive) with them are *X1*, *X2*, *X4* and *X5*. However, this information must be used with caution, as the correlation factor provides a measure of the strength and direction of the linear relationship between two variables and, as it was shown in *Figure 3*, this relationship is far from linear.

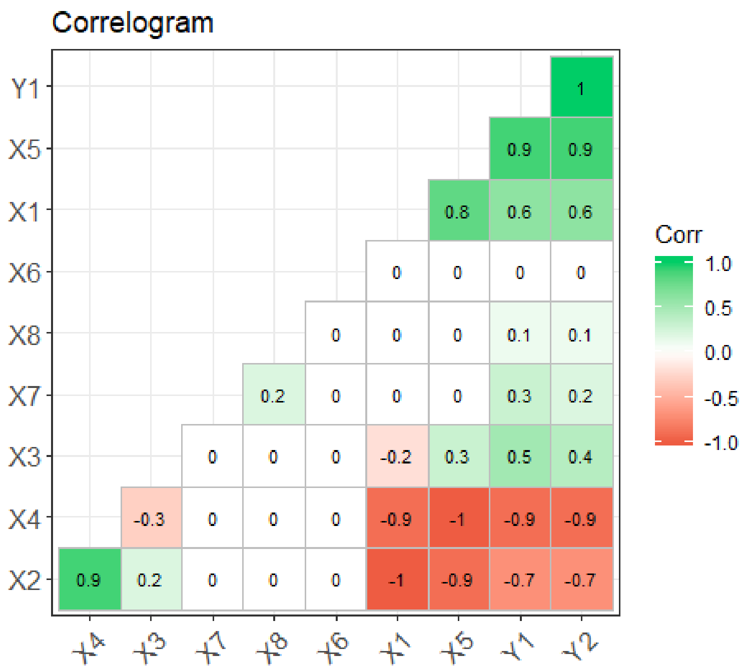


Figure 4 - Correlogram

As a conclusion for the preliminary analysis, we can infer that prediction techniques that assume a normal or Gaussian distribution of the data, as well as linear approaches, will fail to find an accurate relation between the input and the output variables.

Nevertheless, these classical techniques were tried and compared with other, more complex techniques like Random Forest (RF).

2.3 Linear Regression

Given that the output variables span a continuous range of values, using a linear regression technique is the first choice when tasked with predicting the HL and CL for new input variables.

Linear regression is used, as previously mentioned, to stablish the linear relationship between a dependent variable and an independent variable (simple linear regression) or a set of independent variables (multiple linear regression). Since, for our data, there is more than one predictor and two output or dependent variables, two different multiple linear regression models were estimated.

In addition, validation set approach was implemented in order to test the quality of each of those models. The data was divided into two groups, one for training, or building the model, and the other for testing it.

Before starting, however, it is important to mention that multiple linear regression makes several key assumptions [4]:

- Linear relationship between the output variable and the predictors.
- Error terms are uncorrelated.
- Variance of error terms is similar across the values of the predictors (homoscedasticity).
- Predictors are not highly correlated with each other.

These assumptions were tested for each obtained model.

2.3.1 Multiple linear regression

The first thing we tried was to build the regression models for the output variables using all predictors. Upon doing so, we obtained some odd results that led us to believe something was wrong.

After some research, we learned that the problem could be that one of the predictors (*X4*) was perfectly collinear with another one. This was the case, and, in fact, it was not only perfectly collinear, but perfectly multicollinear with another two predictors. The equation they followed was:

$$X4 = \frac{X2}{2} - \frac{X3}{2}$$

To avoid any potential problems caused by this multicollinearity, we decided to remove the predictor *X4* from the data. From this point onwards, *X4* is not considered when building the different models (including the classification models later explained).

Models

The model obtained using the training subset for the heating load (*Y1*) is as follows:

$$Y1 = 78.67 - 61.15\,X1 - 0.08\,X2 + 0.06\,X3 + 4.10\,X5 - 0.01\,X6 + 20.48\,X7 + 0.30\,X8$$

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  78.669755  27.989081   2.811  0.00520 ** 
X1          -61.147782  15.193731  -4.025 6.89e-05 ***
X2           -0.082254   0.025058  -3.283  0.00112 ** 
X3            0.058699   0.009593   6.119 2.35e-09 ***
X5            4.101765   0.492254   8.333 1.46e-15 ***
X6           -0.099626   0.137393  -0.725  0.46883    
X7           20.478418   1.224288  16.727  < 2e-16 ***
X8            0.297998   0.102954   2.894  0.00402 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.047 on 380 degrees of freedom
Multiple R-squared:  0.9088,	Adjusted R-squared:  0.9071 
F-statistic: 540.6 on 7 and 380 DF,  p-value: < 2.2e-16
```

By looking at the p-values for each coefficient, it is possible to see that the predictors that show a stronger linear relationship with the heating load (*Y1*) are *X1*, *X3*, *X5* and *X7*, this last one being the strongest of them all. Likewise, *X6* shows almost no linear relationship with the heating load.

Lastly, a R^2 value of 0.9088 is obtained, meaning that the model fits the training data quite well.

As for the cooling load model, we obtained:

$$Y2 = 91.42 - 68.08\, X1 - 0.08\, X2 + 0.04\, X3 + 4.38\, X5 - 0.04\, X6 + 14.67\, X7 + 0.06\, X8$$

X1, *X3*, *X5* and *X7* still being the variables with a stronger linear relationship with the cooling load, and a R^2 value of 0.8797, which is a bit lower than the heating load case.

Hypothesis check

Now that we have the models, it is important to check if they comply with the four assumptions previously mentioned. Only the model for the heating load will be used to show this conformity or non-conformity with the assumptions (for the cooling load model the results are very similar).

By plotting the residuals versus the predicted values, we obtained the plot shown in *Figure 5*.

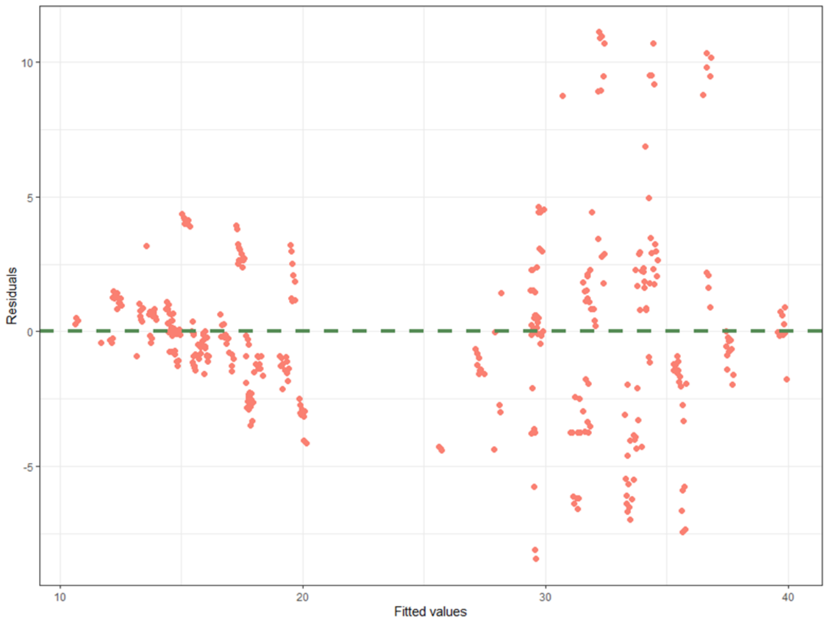


Figure 5 - Residual plot

A funnel shape can be clearly distinguished, which means two things:

1. Errors don't have a constant variance. This is called heteroscedasticity
2. The assumption that there is a linear relationship between the predictors and the output seems to fail. This was expected, as deduced from the scatterplots of *Figure 3*.

To check if the error terms are uncorrelated, a plot with the residuals as a function of time can be made. This plot is shown in *Figure 6*.

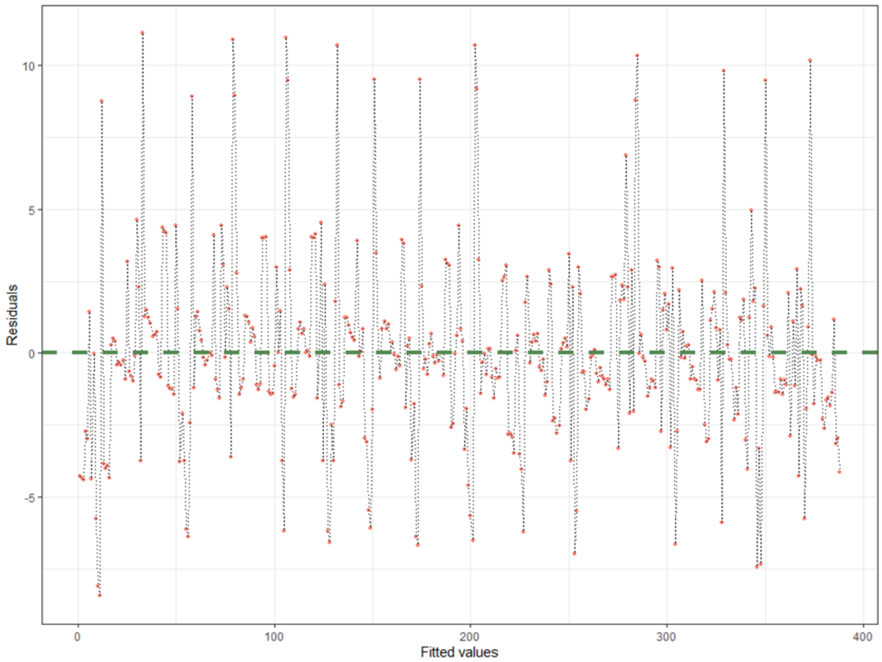


Figure 6 - Correlated errors plot

A pattern can be discerned, and a lot of adjacent residual terms are on the same side of the plot, which means the error terms are correlated.

Lastly, the correlation between predictors can be checked with the Variance inflation factor (VIF). If said factor exceed 10 then a problematic amount of collinearity is present between the predictors.

We obtained a VIF of 47.9 for both the heating and cooling load models.

As was expected, the linear models present a lot of problems when fitted with data with the statistical properties described in the preliminary analysis

Mean Square Error (MSE)

The Mean Square error (MSE) is the average squared difference between the estimated value and the actual value for the test group. The formula for MSE is shown below:

$$MSE_{VS} = \frac{1}{t} \sum_{i \in \Omega_{TEST}} (y_i - \hat{y}_i)^2$$

We obtained an MSE value of 8.13 for the heating load and a value of 9.66 for the cooling load, which are quite high.

Best subset selection

Even though in this case the number of predictors is low, it is often a good idea to improve the model by removing irrelevant variables, thus enhancing its interpretability.

Then again, as there aren't a lot of variables, all the possible models resulting of the combinations of said variables can be tested, and then, using parameters such as C_p, BIC or Adjusted R^2 the best model can be chosen among them.

After applying the exhaustive selection algorithm, the following results were obtained for the heating load (left) and cooling load (right) models. Also, the plots for the parameters C_p (left), BIC (center), Adj R^2 (right) can be seen on the top row of *Figure 7* for the heating load and on the bottom row for the cooling load.

```
         X1  X2  X3  X5  X6  X7  X8
1  ( 1 ) " " " " " " "*" " " " " " "
2  ( 1 ) " " " " " " "*" " " "*" " "
3  ( 1 ) " " " " "*" "*" " " "*" " "
4  ( 1 ) "*" " " "*" "*" " " "*" " "
5  ( 1 ) "*" "*" "*" "*" " " "*" " "
6  ( 1 ) "*" "*" "*" "*" " " "*" "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
```

```
         X1  X2  X3  X5  X6  X7  X8
1  ( 1 ) " " " " " " "*" " " " " " "
2  ( 1 ) " " " " " " "*" " " "*" " "
3  ( 1 ) "*" " " " " "*" " " "*" " "
4  ( 1 ) "*" " " "*" "*" " " "*" " "
5  ( 1 ) "*" "*" "*" "*" " " "*" " "
6  ( 1 ) "*" "*" "*" "*" "*" "*" " "
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
```

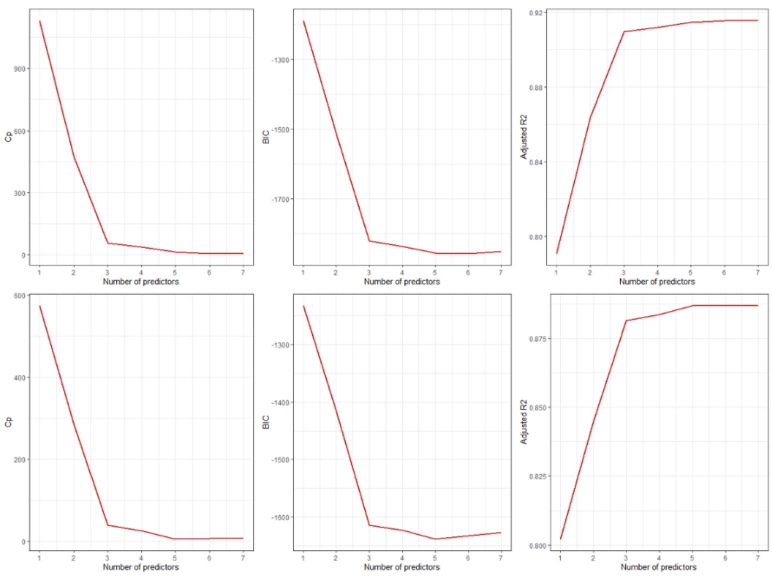


Figure 7 - Subset selection parameters for HL (top) and CL (bottom)

Heating load	
Parameter	Best Model
Cp	6
BIC	6
Adj R2	6

Cooling load	
Parameter	Best Model
Cp	5
BIC	5
Adj R2	6

For the heating load the best model is the one with six predictors (excludes *X6*), and for the cooling load the one with five predictors (excludes *X6* and *X8*).

2.4 Classification

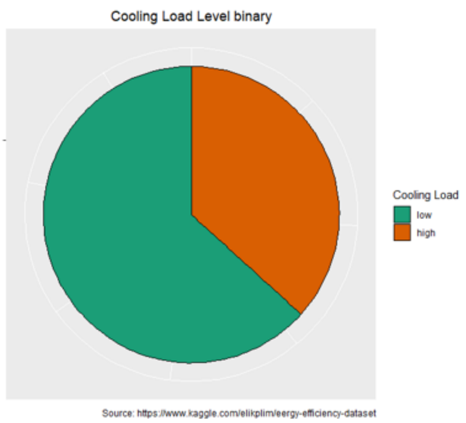
Statistical Classification describes the process of assigning a category to a new observation, based on a training set where the category of each observation is known. The category can be multiclass or binary. The implemented algorithm that in the end transacts the classification is called classifier. There are many different models for classification, but we focused on:

- Linear Discriminant Analysis (LDA)
- k-nearest Neighbors
- Classification Trees
- Random Forest

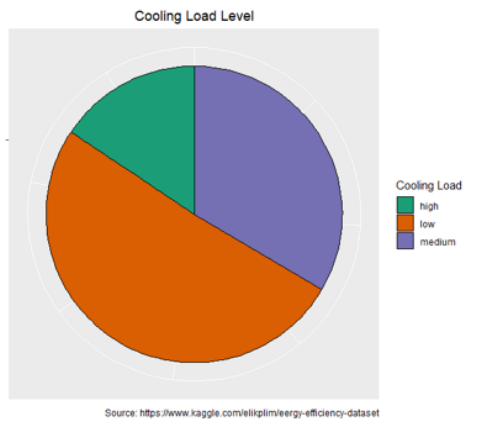
Looking at the data, the use of these tools may initially seem counterintuitive. However, in practice it can be convenient to discretize the output variable and treat the resulting application as a binary or multiclass classification problem.

In order to do this, the output *Y2* for the cooling load was categorized into "high", "medium" and "low". Therefore, the range of the variable was considered and afterwards divided into three equal sections. That resulted into the distribution showed in the right pie chart and table of *Figure 8*.

Where some models can work with multiclass or binary outputs, others just operate with binary output variables. To be able to compare between binary and multiclass classifier, the output cooling load was additionally transferred into binary, by setting a cutoff in half of the range (*Figure 8*, left)



Low	High
486	282



Low	Med	High
391	257	120

Figure 8 - Binary and multiclass distribution for variable Y2

2.4.1 Linear Discriminant Analysis (LDA)

The first classifier used is the LDA. It is used to find a linear combination of predictors to categorize two or more classes. For using the LDA the following assumptions need to be fulfilled:

- Common covariance across all response classes σ^2
- Distribution of observation in each of the response classes is normal with a class-specific mean μk and common covariance σ. [5]

As mentioned before, the predictors used in the dataset are not distributed normally and their relationship with the output variables is not linear. That is why it is expected that the LDA is not an adequate classification model for this dataset.

The results show, that the training error is way higher than in the other approaches. The LDA is not considered further.

	High	Low	Med
High	64	0	56
Low	0	384	7
Med	44	0	213

$$Train\_Error = \frac{56 + 7 + 44}{768} = 0.139$$

2.4.2 k-nearest Neighbors (KNN)

KNN is a classifier based on Probabilities. For a new observation it at first searches for k observations in the training data that are closest to the new observation. Then the classifier calculates the conditional probabilities for each class. At the end it classifies the new observation as the class with the highest probability.

For the analysis of the before mentioned binary dataset, at first the number of neighbors was set to $k = 10$. The results obtained are relate to the use of the whole set as training and test.

	Low	High
Low	448	38
High	19	263

The confusion matrix shows how well the classifier performs. Out of “false positives" and "false negatives" we can calculate the error of the prediction, where CM refers to the confusion matrix on the left:

$$Error = \frac{CM(1,2) + CM(2,1)}{Sum(CM)} = \frac{57}{768} = 0.074$$

k-fold cross-validation

Cross-validation is used to judge the quality of a model. The idea is at first, to separate the model randomly into *k* parts. Afterwards one of the parts will be used to test the model and other *k-1* parts are used to train the model. Then the test error is calculated as usual. In the next cycle another part will be used as test set and again the model will be trained with the rest and the test error calculated. At the end the model will have been trained and tested *k* times and *k* test errors will be calculated. The mean of these errors is the result of the k-fold cross-validation.

The binary set will be divided into five parts. The procedure can be implemented quite elegantly in R by assigning a number between $1\!:\!k$ randomly to each row of the set. In a loop it is possible to execute the whole calculations without a lot of effort.

```
for(i in 1:k){

  knn.pred = knn(data.knn.X[folds!=i,],data.knn.X[folds==i,],data.knn.Y[folds!=i],K)

  fold.t =table(engy_knn$load_bin[folds==i],knn.pred)
  cv.error[i] = (fold.t[1,2] + fold.t[2,1])/sum(fold.t)
}
```

The function *knn()* executes training and test in one single operation. With the *folds* the parts are addressed. In the next step the errors will be calculated and saved in a vector that in the end contains all the *k* errors.

As a result, for the k-fold cross-validation with $k = 8$ folds, the following test errors were obtained:

```
> cv.error
[1] 0.07954545 0.08791209 0.09473684 0.12903226 0.09638554 0.06422018 0.05813953 0.08943089
> mean.cv.error
[1] 0.08742535
```

2.4.3 Classification Trees

Classification or decision trees are a very intuitive method for classifying multiclass or binary outputs. Trees mainly consist out of nodes and branches. In every branch, it will be decided binarily which way to take, depending on the value of the associated predictor.

The tree shown on *Figure 9* describes the predictions for the multiclass dataset.

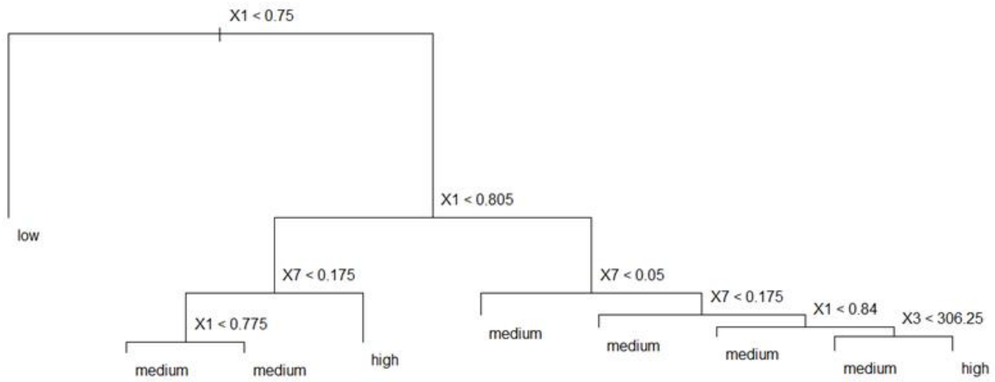


Figure 9 - Multiclass classification tree

In the classification trees it is very easy to see which variables have a lot of impact in the model. Like this, it is possible to figure out that for predicting a low cooling load, it is just necessary to consider predictor $X1$. To distinguish between medium and high $X1$, $X7$ and $X3$ are used. It is possible that the same predictor is used in multiple branches with different thresholds.

The tree for the binary output variable looks quite similar, while this one just uses two predictors for the classification. It is shown on *Figure 10*.

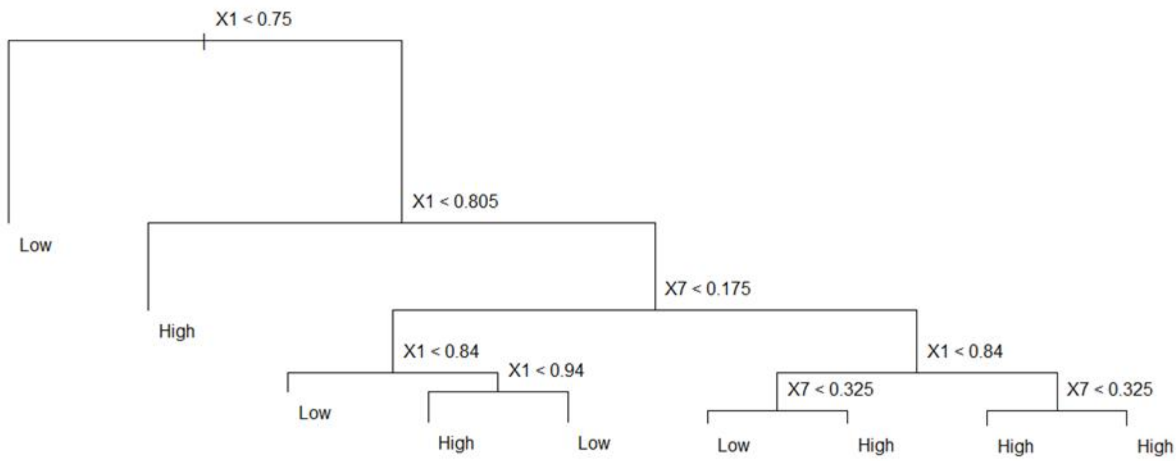


Figure 10 - Binary classifaction tree

2.4.4 Random Forest

This model is a collection of afore mentioned classification trees to improve performance. Each time building a tree, another training set will be used. Due to this procedure, we can obtain the importance of the predictors as a byproduct.

Each cut just considers a small set of explanatory variables, usually $\sqrt{k}$ predictors. In this case with seven predictors the number of predictors used in every tree can be set to $mtry = 3$.

	High	Low	Med	Cl. Er.
High	91	0	29	0.242
Low	0	388	3	0.008
Med	16	2	239	0.070

On the right column the class error is displayed. The highest error appears for the class "high". This could be explained because this class has the least observations. Summing up the false predictions and dividing by the total number of observations gives the total $error = 0.0651$.

As mentioned before, with random forest it is possible to plot the importance of variables as well. *Figure 11* show two different approaches.

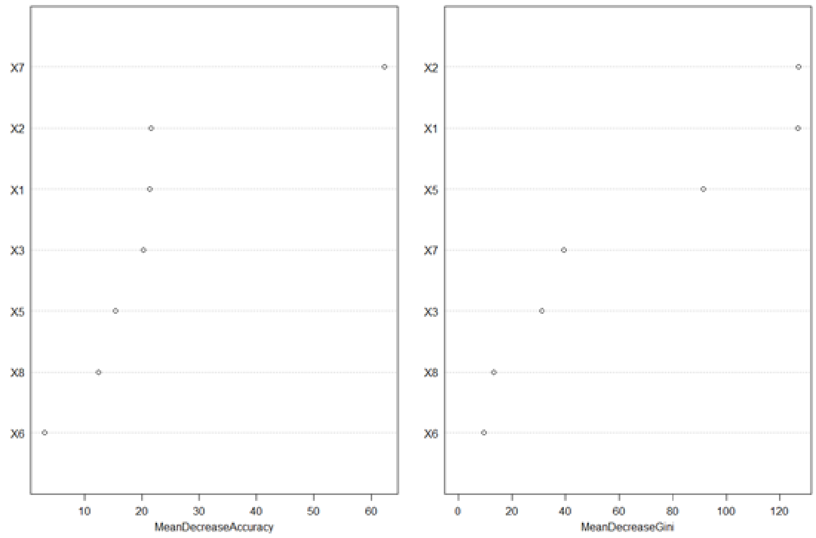


Figure 11 - Random Forest variable importance

The Mean Decrease Gini is based on Gini Impurity, which is used in classification trees to decide how to split the data into smaller groups. Mean Decrease Gini is the average mean of a predictor's total decrease in node impurity.

The Mean Decrease Accuracy focuses on the errors, for each tree, the prediction error on the out-of-bag portion of the data is recorded. Then the same is done after permuting each predictor variable. The difference between the two are then averaged over all trees and normalized by the standard deviation of the differences.[6]

The two approaches do not bring exactly the same results for importance of variables, due to the different points of view. Still it is possible to extract that $X1, X2$ and $X7$ have high importance for the model.

2.5 Principal Component Analysis

As an addendum, the Principal Component Analysis of the data has also been carried out, even though prediction models were not built or tested from the obtained results. The collinear predictor, *X4*, has also been taken into account for this analysis, as the PCA is really useful when a set of variables are highly correlated.

We obtained the following results:

- $Z1 = 0.496X1 - 0.502X2 - 0.505X4 + 0.496X5$
- $Z2 = 0.245X1 - 0.232X2 - 0.894X3 + 0.206X4 - 0.210X5$
- $Z3 = -0.707X7 - 0.707X8$
- $Z4 = X6$
- $Z5 = -0.707X7 + 0.707X8$
- $Z6 = 0.495X1 + 0.291X3 - 0.205X4 - 0.790X5$
- $Z7 = 0.670X1 + 0.505X2 + 0.450X4 + 0.293X5$
- $Z8 = -0.660X2 + 0.327X3 + 0.677X4$

Where *Z1* explains 46% of the variance in the original variables, *Z2* explains 16%, *Z3* explains 15%, *Z4* explains 12%, *Z5* explains 10%, *Z6* explains 1% and both *Z7* and *Z8* explain almost 0%.

Therefore, *Z1*, *Z2*, *Z3* and *Z4* account for 89% of the original information.

3 Conclusions

At the beginning of the report, we said that it was our objective to answer the question of how the shape and orientation of a building affects its energy efficiency, or rather two variables tightly related with it. For this purpose, we tried different techniques, obtaining different results for each of them.

As we detected in the preliminary analysis, the complexity of the data makes that its statistical properties are not the most appropriate to be approached with the more classical techniques, such as linear regression. This was later proved true, as the obtained multiple linear regression models for both the heating load and cooling load presented many problems that diminish its validity and suggest caution when using them to predict new observations. These problems, collinearity in particular, could be solved, in part, by using the variables obtained in the PCA, but it was not tried in this report.

However, if we had to answer the afore mentioned question based on the multiple linear regression models, we could say that, apparently, the cooling load is harder to predict than the heating load (lower R^2 values and higher errors). The most important predictor seems to be *X7*, which is the glazing area, even though it is not the most correlated with either the HL or CL. This makes sense, as the amount of glass area present in a building determines the heat absorbed due to the sun, but it also is the source of heat leakage from the building to the environment. And, the predictor *X6*, which is the orientation, seems not to matter in the slightest.

Then, after categorizing one of the output variables (*Y2*), some other classification techniques were used to try to answer the question. All these techniques yielded pretty good results with low error rates, excluding LDA, which obviously had the same problem as the linear regression approach.

It is true, however, that the predicted category englobes a multitude of values, so following the approach of a binary or multiclass (3) classification will not allow to get an exact estimation of the cooling or heating load. Though, greatly increasing the number of categories could solve this problem, and the more modern and complex classifications tools can handle it well.

Nevertheless, one thing that the classification approach, specifically Random Forest, seemed to confirm, is that the most important factors to take into account to determine the HL and CL, and thus the energy efficiency of a building, are the glazing area, the surface area and the relative compactness, while also considering of little importance the orientation.

Finally, to sum it all up. We have tried different techniques with mixed success, but, at the end of the day, we consider that the proposed question has been, at least, partially answered. Hopefully, further analysis, better data adaptation and the usage of more complex tools can help obtain more exact and faithful results.

4 References

[1] https://ec.europa.eu/energy/en/topics/energy-efficiency/energy-performance-of-buildings/overview

[2] https://www.basix.nsw.gov.au/iframe/basix-help-notes/thermal/heating-and-cooling-loads.html

[3] http://people.maths.ox.ac.uk/tsanas/Preprints/ENB2012.pdf

[4] https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/

[5] https://datascienceplus.com/how-to-perform-logistic-regression-lda-qda-in-r/

[6] https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/importance