Danmarks Tekniske Universitet

DTU

02450 Project 1

Authors

Group 192:
Amine Lbath - s221881
Eirik Runde Barlaug - s221409
Sara Bakken Heiberg - s221630

Student	Section 1	Section 2	Section 3	Section 4	Exam questions
s221881	40%	30%	30%	33.33%	33.33%
s221409	30%	40%	30%	33.33%	33.33%
s221630	30%	30%	40%	33.33%	33.33%

Table 1: Contribution of each member

October 25, 2022

Contents

1 Introduction

1.1 Presentation and objectives

With global warming and the energy crisis of 2022, energy efficiency has become more crucial than ever. According to the U.S. Department of Energy, heating and air conditioning account for over 40 % of households' energy bills in 2021. Therefore, it seems essential to design more energy efficent buildings in the future. In recent years, machine learning techniques have been applied to analyze and predict energy efficiency of buildings. This can be used to forecast and minimize the energy consumption, while maintaining or even enhancing the indoor environmental quality. To help with that, it is possible to predict the future energy consumption with data describing the design of the building. The Energy Efficiency data set [1] was created for this specific purpose.

The data set shows the result of energy analysis using 12 different building shapes simulated in Ecotect. Ectoect is an environmental analysis tool used to simulate builing performance, including energy consumption.

The data set comprises 768 samples (i.e. unique buildings), obtained by simulation with various settings of 8 different building parameters: the relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area and glazing area distribution. The simulation outputs two real valued responses that are good indicators for building energy efficiency, namely the heating load and the cooling load. These two are the variables we will treat as our responses.

In this project our main focus will be vizualisation and description of the data, as well as a part regarding dimensionality reduction. This project will act as the basis for another project where we will eventually try to predict and classify the heating and cooling load using multiple linear regression as well as multiclass classification. For the latter, the responses must be discretized into classes, which is why both the heating and cooling load will be discretized into four classes each in this project.

1.2 Previous analysis of the data

The 2012 paper *Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools* [2] written by Athanasios Tsanas and Angeliki Xifara used the energy efficiency data set. At the time of the article, the most popular method to compute the heating and cooling load for efficient building design, was to use various simulation software. This is quite time-consuming and can only be handled by an expert in the particular software.

The goal of the paper was therefore to provide a faster approach to predictive analysis of building energy consumption. They used both the Iteratively Reweighted Least Squares method and the Random Forests method. The conclusion was that Random Forests was an efficient method to predict HL and CL. This approach computes a sufficiently accurate prediction in a matter of seconds.

They also noted that the relative compactness, wall area and roof area appeared to be mostly associated with HL and CL. Furthermore, the glazing area seemed to be the most

significant variable to estimate the energy performance of a given building. Finally, they observed that HL can be estimated more accurately than CL with their model and the current data set.

2 Detailed explanation of the attributes of the data

In this section, we will give a detailed explanation of the attributes of the data set. That is, we will describe the data, its physical interpretation and present a statistical summary of all attributes.

2.1 Attribute description

Table 2 summarizes some key properties of the 10 attributes in the energy efficiency data set.

Type	Attribute	Description	Datatype	Unit
Feature	X_1	Relative compactness	Continuous, Ratio	None
Feature	X_2	Surface Area	Continuous, Ratio	m^2
Feature	X_3	Wall area	Continuous, Ratio	m^2
Feature	X_4	Roof area	Continuous, Ratio	m^2
Feature	X_5	Overall height	Continuous, Ratio	m
Feature	X_6	Orientation	Discrete, Nominal	None
Feature	X_7	Glazing area	Continuous, Ratio	None
Feature	X_8	Glazing area distribution	Discrete, Nominal	None
Response	Y_1	Heating load	Continuous, Ratio	kWh/m^2
Response	Y_2	Cooling load	Continuous, Ratio	kWh/m^2

Table 2: Summary of the variables in the energy efficiency data set.

Conveniently, the data set does not contain any missing values and is not corrupted in any way. *It is however important that the reader is aware of the following:* The continuous variables in the data set are continuous in the sense that they can take on any real value. Nonetheless, since the data set is constructed in such a way that the researchers have chosen some specific values for all of the continuous variables, and simulated buildings based on the combinations of those, it may appear "discrete" in some plots. This is just because some points lie on top of each other.

2.1.1 Physical interpretation of the attributes

- **X1 - Relative compactness:** Measures the compactness of the building. I.e., the buildings volume to surface ratio compared to the most compact shape with the same volume.[3] An example of different building shapes and their relative compactness is given in figure 1.

- **X2, X3, X4, X5 - Surface area, Wall area, Roof area, Overall height:** Inside surfaces that would need heating/cooling, as well as the height of the builing.
- **X6 - Orientation**: The major orientation of the building (2 = North, 3 = South, 4 = East, 5 = West).
- **X7 - Glazing area:** The proportion of "glazing" (i.e. windows, glass walls etc.) relative to the floor surface.
- **X8 - Glazing area distribution:** The distribution of the glazing on the builing. This feature is categorized into five different distributions. 1 = Uniform distribution, i.e. 25% of the glazing in each cardinal direction. 2, 3, 4, 5 means that 55% of the glazing lies on the corresponding wall from the orientation feature, and 15% is placed in the other three cardinal directions.
- **Y1, Y2 - Heating- and cooling load:** The amount of energy needed to heat or cool the building. Given in kilowatt-hours per meters squared.

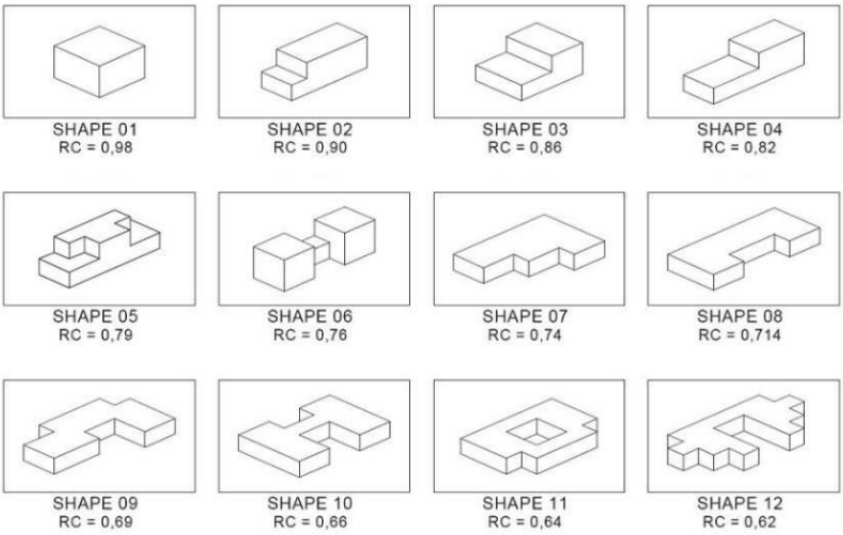

Figure 1: Examples of different buliding shapes' relative compact ratio (RC). Source: [3]

2.2 Summary statistics of the data

In this section the basic statistical properties of the data set is described. Table 3 below summarizes these statistics for all attributes:

Attribute	Count	Mean	Std.dev.	Min	Q1	Q2	Q3	Max
X1	768	0.764	0.106	0.620	0.683	0.750	0.830	0.980
X2	768	671.708	88.086	514.500	606.375	673.750	741.125	808.500
X3	768	318.500	43.627	245.000	294.000	318.500	343.000	416.500
X4	768	176.604	45.166	110.2500	140.875	183.750	220.500	220.500
X5	768	5.250	1.751	3.500	3.500	5.250	7.000	7.000
X6	768	3.500	1.119	2.000	2.750	3.500	4.240	5.000
X7	768	0.234	0.133	0.000	0.100	0.250	0.400	0.400
X8	768	2.813	1.551	0.000	1.750	3.000	4.000	5.000
Y1	768	22.307	10.090	6.010	12.993	18.950	31.668	43.100
Y2	768	24.588	9.513	10.900	15.620	22.080	33.133	48.030

Table 3: Summary statistics of the attributes.

3 Data visualization

In this section, we are going to visualize the data set in order to enhance our understanding of it. In particular, we will analyze the distribution of the data to try to get a sense of the relations between the different attributes

3.1 Distribution of the attributes

Figure 2 and 3 describes the distribution of the attributes in terms of a boxplot, a histogram and the kernel density, respectively.

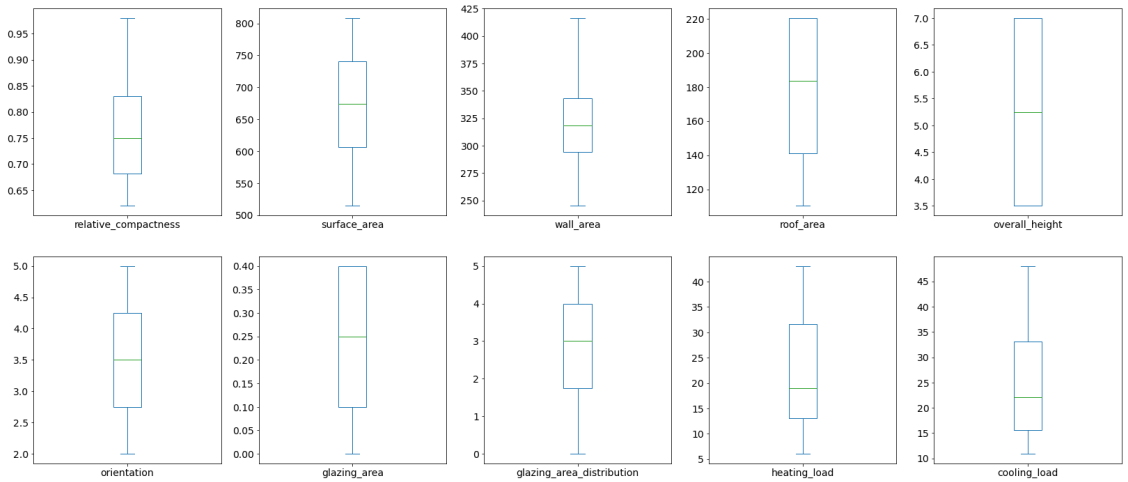

Figure 2: Boxplots of the attributes.

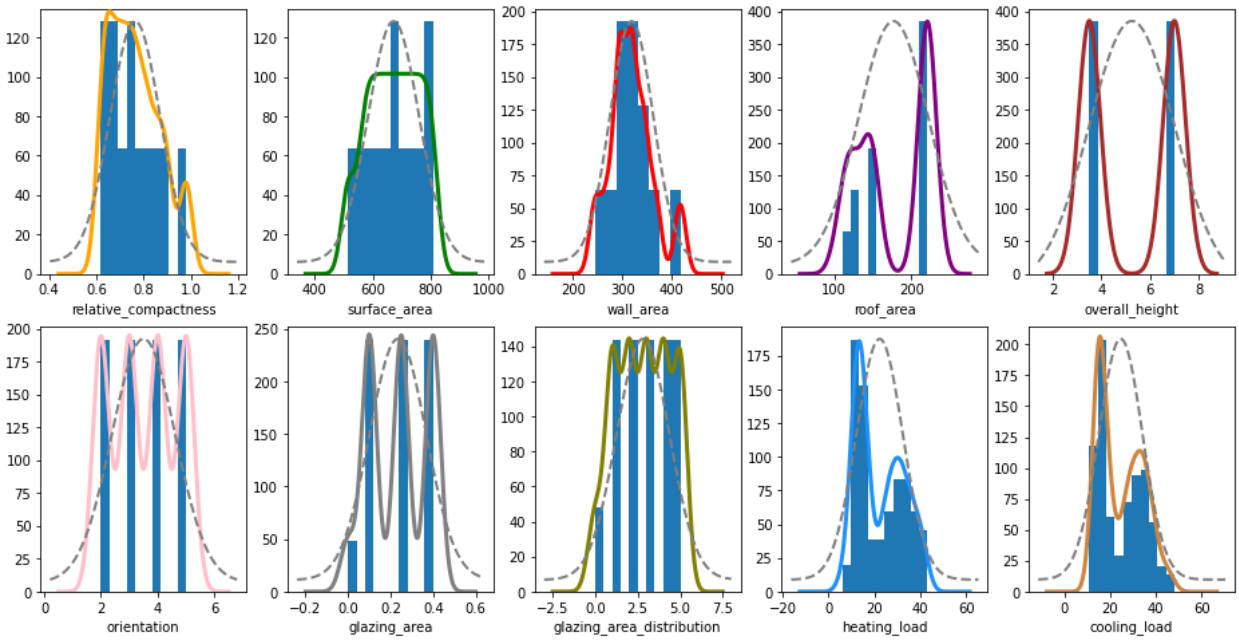

Figure 3: Histogram, Kernel density estimation and Normal distribution for each attribute.

As one can observe in both figures above, there does not seem to be any extreme values. What's more, the values are consistent with the physical quantities they represent. Thus, we can ensure that there are no outliers within the data set.

Figure 3 above, also reveals that only the relative compactness, the surface area and the wall areas tend to follow a normal distribution. The other variables are either nominal and follow a normal distribution or they follow totally different distributions.

3.2 Relation between the features and the responses

In the following we will try to get a visual overview of dependencies between the features (the builing parameters) and the responses (the heating and cooling load). First, we will as earlier mentioned discretize the heating and cooling load, so as to visualize these attributes better in a pair plot. Based on the quantiles of our data and using the histogram of the loads, we divided the data into four classes by the following regime:

Given some data point $\mathbf{x}$, its heating load $HL(\mathbf{x})$ and cooling load $CL(\mathbf{x})$

$$\begin{aligned}
\mathbf{x} \in \text{Class } 1 &\iff 0 \leq HL(\mathbf{x}) \leq 10 \quad \vee \quad 0 \leq CL(\mathbf{x}) \leq 20 \\
\mathbf{x} \in \text{Class } 2 &\iff 10 < HL(\mathbf{x}) \leq 20 \quad \vee \quad 20 < CL(\mathbf{x}) \leq 30 \\
\mathbf{x} \in \text{Class } 3 &\iff 20 < HL(\mathbf{x}) \leq 30 \quad \vee \quad 30 < CL(\mathbf{x}) \leq 40 \\
\mathbf{x} \in \text{Class } 4 &\iff HL(\mathbf{x}) \geq 30 \quad \vee \quad CL(\mathbf{x}) \geq 40
\end{aligned}$$

The pair plots in figure 4 and figure 10 displays the relation between the attributes.

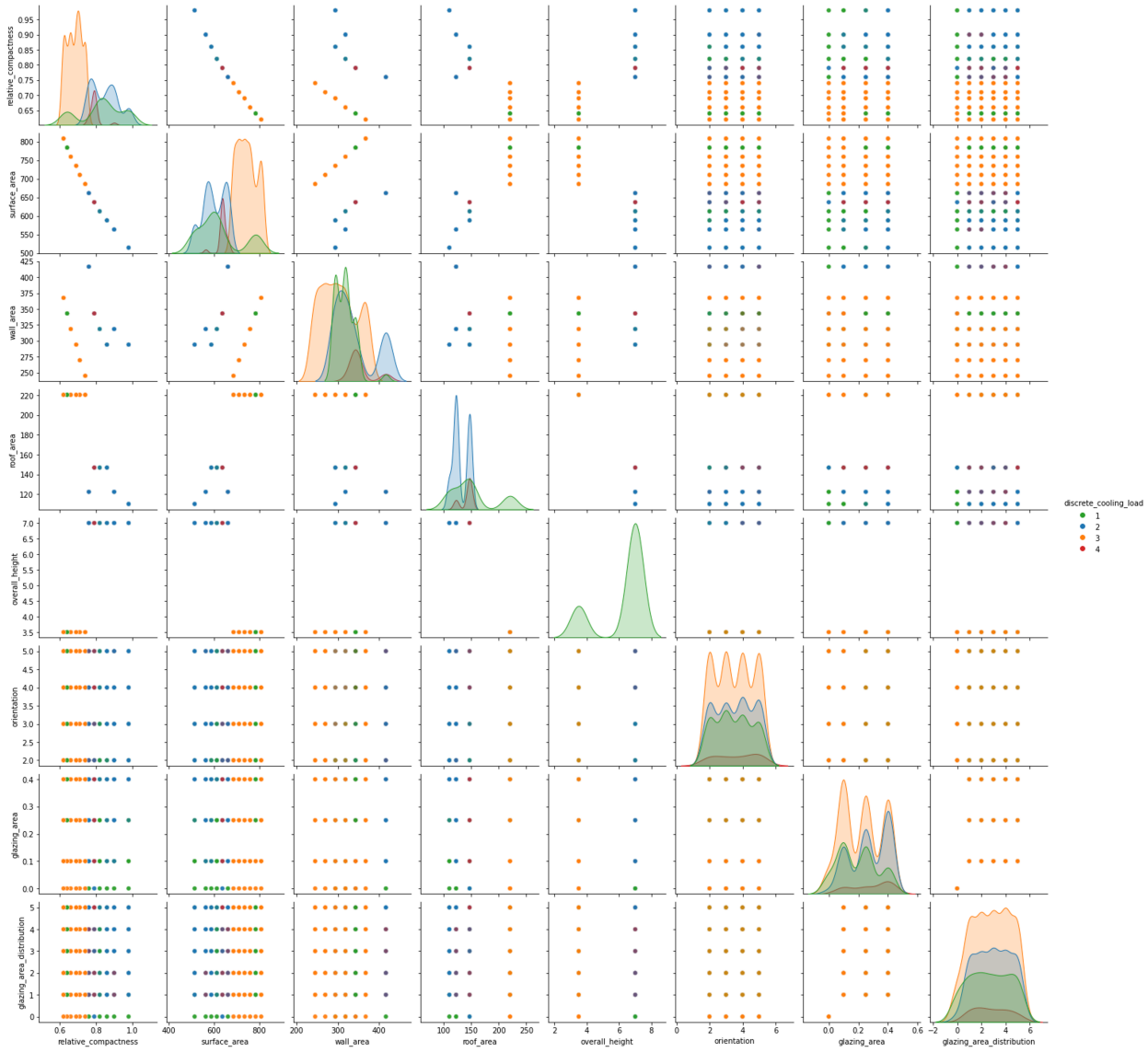

Figure 4: Pair plot for the cooling load

In the above plot, we can see that some pairs of attributes tend to discriminate *high* and *low* loads better than others. For instance, smaller relative compactness and overall height generally lead to high cooling load and low heating load and vice versa. On the other hand, it seems harder to discriminate loads using orientation and glazing area distribution. All different load values does indeed overlap on almost all possible values.

To get an even better understanding of the link between attributes and loads, we will now analyze the relation between each attribute and the continuous heating and cooling loads.

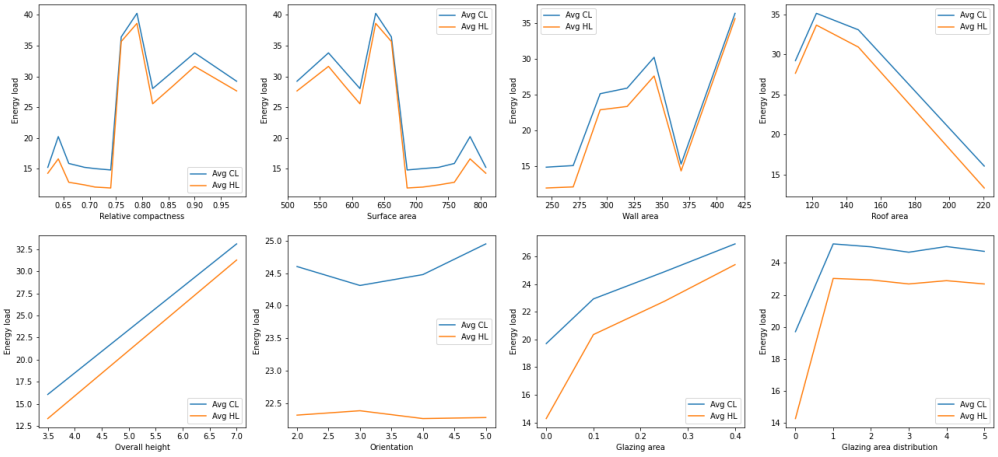

Figure 5: Average heating and cooling load for every attribute

The plots above show that there are indeed a couple more relations between loads and attributes. For instance, smaller relative compactness corresponds to lower loads and vice versa. Besides, the overall height seems to be proportional to the load. These relations are further examined in chapter 3.3.

3.3 Correlation

The correlation between the different attributes is given in the correlation matrix in figure 6.

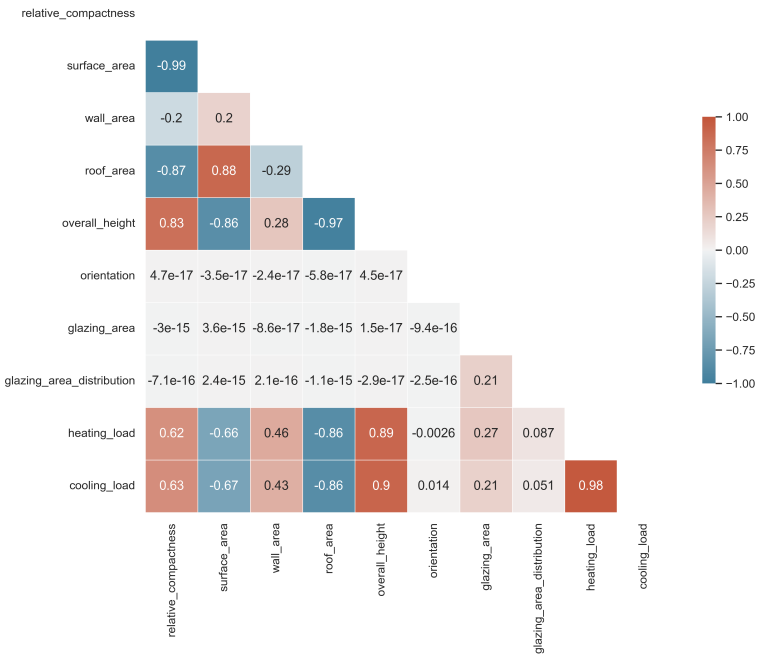

Figure 6: Correlation matrix of the attributes.

The correlation matrix confirms, and quantifies, some of the earlier mentioned relations between the attributes. As one can see, orientation and glazing distribution does not seem to correlate significantly with any of the other attributes. Some of the correlations between the features are present due to the shape of the simulated buildings. E.g., a building with a high surface area does also, intuitively, have a low relative compactness. However, it is interesting to observe how the heating- and cooling loads are highly correlated with each other, as well as some of the other features. It does for instance seem like higher values for height, wall area and relative compactness yields high heating- and cooling loads. On the contrary, surface and roof area seem to have the opposite effect on the two responses.

3.4 Principal Component Analysis

The features of the data resides in a 8 dimensional space. As this will make the visualization of the data difficult and also require a lot of storage space, we use principal component analysis to transform the data into fewer dimensions, while preserving the most important information.

Figure 7 shows the percent of variance explained by each of the principal components. It also displays the accumulation of variance for each principal component. This shows that almost 90% of the variance in the data is explained by the first four principal components.

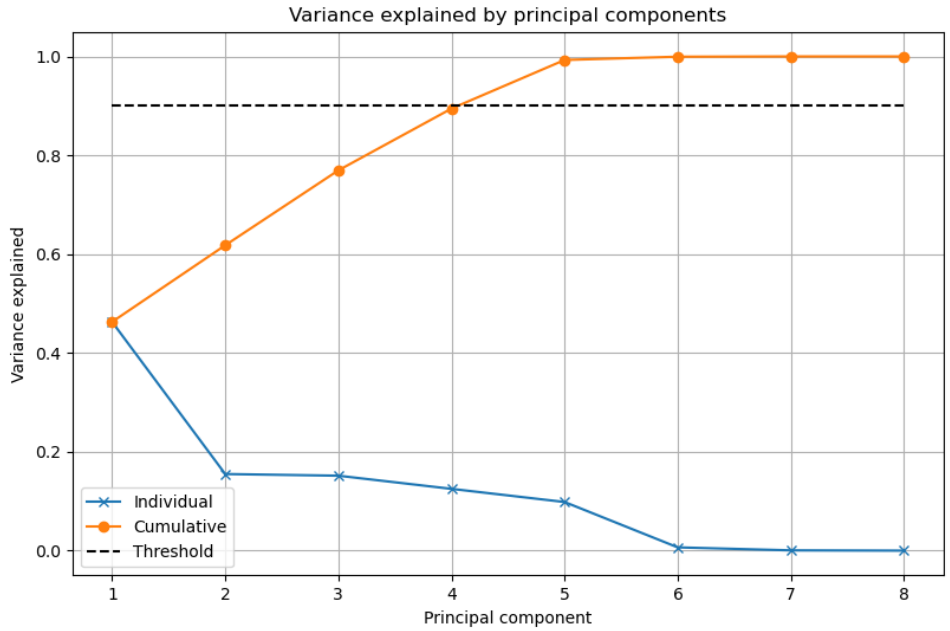

Figure 7: Variance explianed by principal components

From figure 8 we can see that five out of eight attributes are mainly described by the first two principal components. This is reasonable as figure 7 shows that approximately 62,5% of the variance is explained by the first two principal components (PC1 and PC2).

The last two attributes are described by the two last principal components (PC3 and PC4). This makes sense as figure 7 shows that 90% of the variance is explained by the first four principal components.

Consequently, a projection onto a subspace of minimum 4 dimensions is necessary in order to preserve sufficient information about the data set.

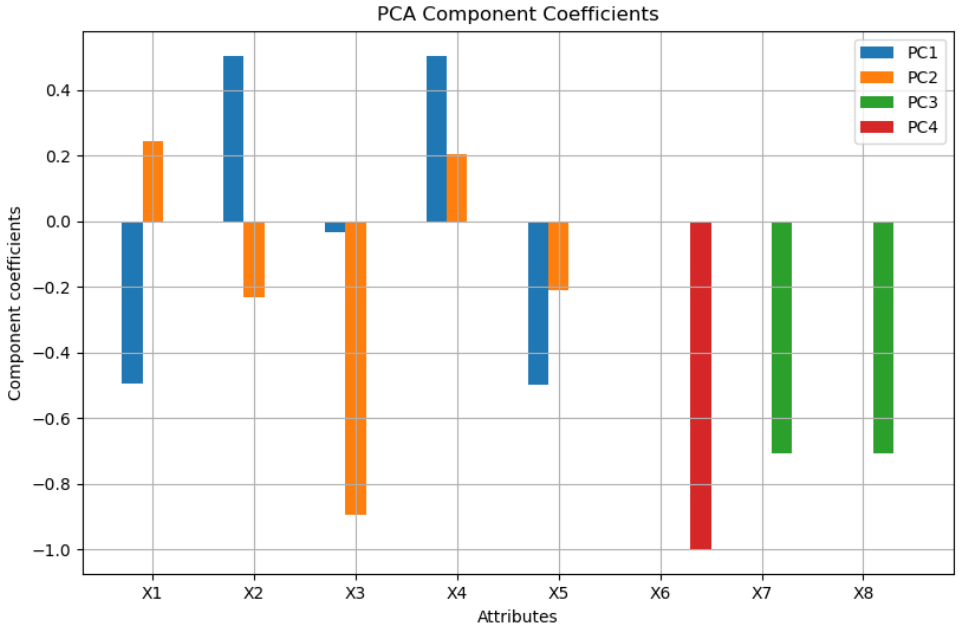

Figure 8: PCA component coefficients

By projecting all the observations onto the 4-dimensional subspace and plotting the results, it is possible to visualize the structure of the transformed data set. The projection of the data on the PCA-space creates a visualization that minimizes residual variance that cannot be explained by the variables in the model (i.e. noise), and maximizes the variance of the projection coordinates. As it is hard to visualize a 4-dimensional space, we have plotted the projected data onto pair plots of principal components in figure 9. For some pairs of PCs, this results in a good separation of the classes of the cooling load.

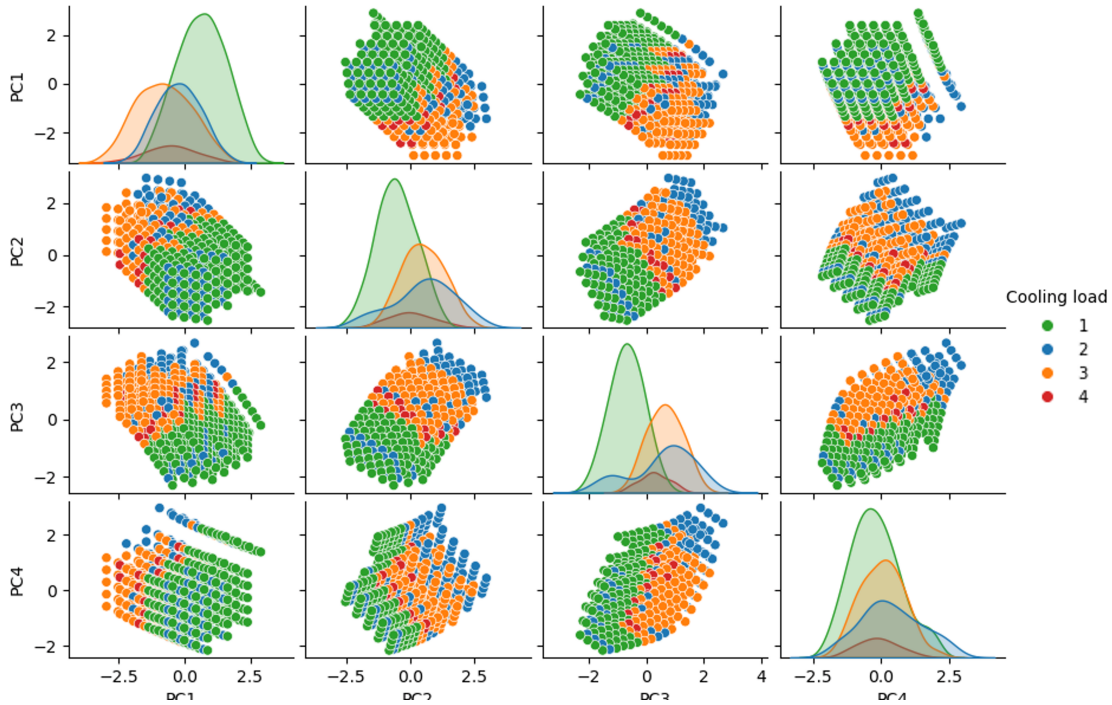

Figure 9: Projection

4 Key lessons about the data

The energy efficiency data set is composed of 8 major features representing building parameters, with two associated responses: Heating and cooling loads. Our main goal consists of modeling the energy loads using building features.

As the data set was created by researchers from scratch using a modeling software, it does not contain any missing values or outliers. Hence, no data preparation is required, which makes it really effortless to work with and well suited for the task ahead.

Yet, it is important to note that not all attributes follow a normal distribution. While some might follow normal or uniform distributions, others do not follow any common distribution.

Furthermore, we have found some major relationships between the attributes and the responses. This has been investigated both visually through plots and quantitatively by the correlation between the attributes.

By using principal component analysis we reduced the dimensionality of the data set. Also, projecting our data on the PCA subspace, visually confirmed that models could be derived to predict the energy load

All in all, it seems as if the primary machine learning aim does indeed seem feasible.

5 Appendix A: Heating load pair plot

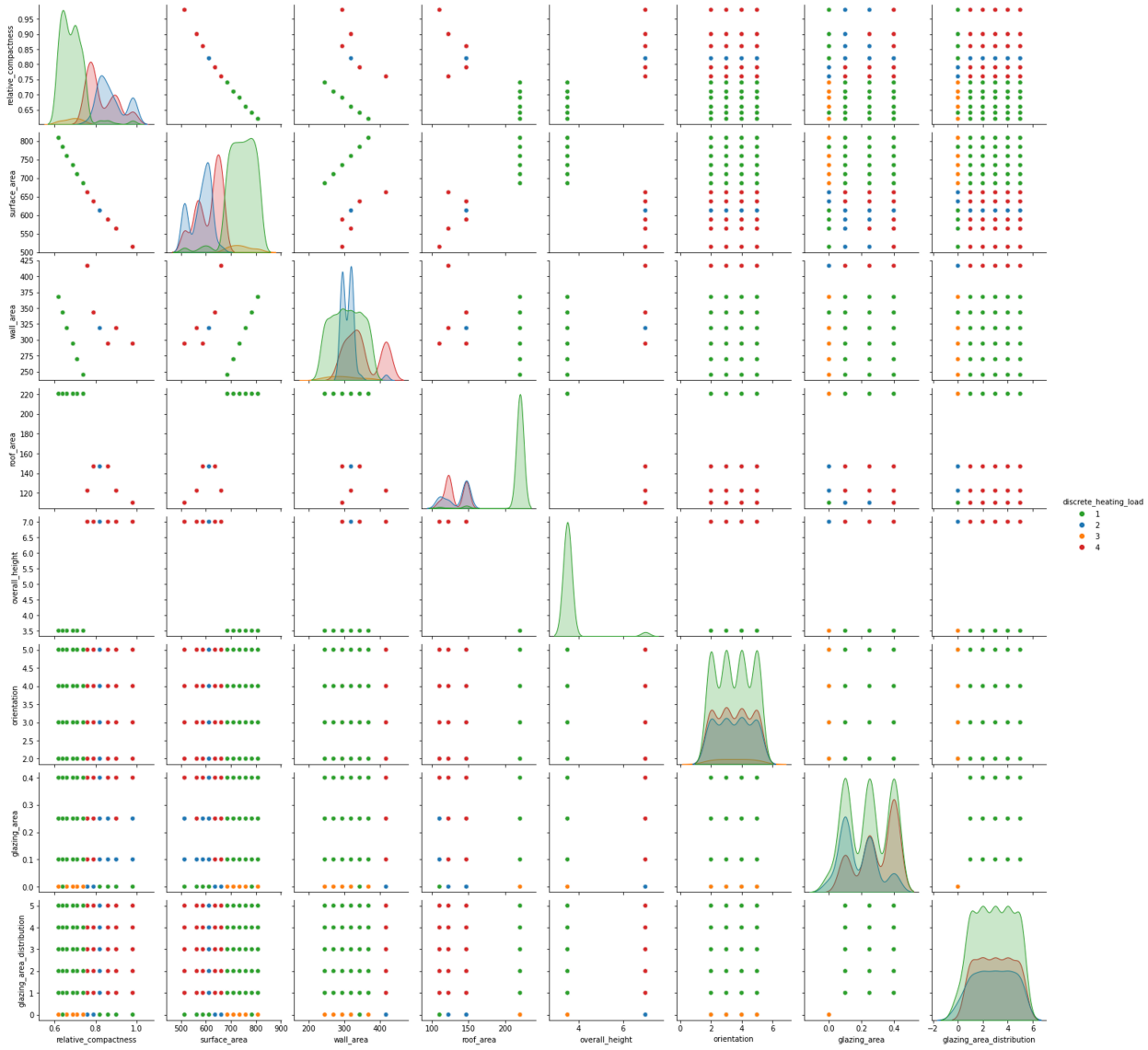

Figure 10: Pair plot for the heating load

6 Appendix B: Exam problems

1. **C**: *Time of day* and *Congestion level* are discrete categories which can be ranked, hence they are ordinal variables. The *Traffic lights* and *Running over* attributes are counters, hence they are ratio.

2. **A**: max{|19-26|, |0-2|} = 7

3. **A**: $\frac{\sum_{i=1}^{4}\sigma_i^2}{\sum_{i=1}^{5}\sigma_i^2} = 0.87 > 0.8$

4. **D**: The values related to **Broken Truck**, **Accident victim**, and **Defects** in the the second column of V are positive and non-negligible. On the other hand, the value related to **Time of day** is negative. The value associated to *Immobilized bus* is very small and positive. Hence, taking a high value for the positive attributes and a small value for the negative one, will result into a positive value of the projection onto principal component number 2.

5. **A**: Since we are using bag-of-words encoding, each unique word in document i is considered as an element of the set s_i. Thus we have that $J = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|} = \frac{\#\ Unique\ words\ in\ both}{\#\ Unique\ words\ in\ total} = \frac{2}{13} = 0.153846$

6. **B**: The probability that an observation had $\hat{x}_2 = 0$ given that the congestion level is light ($y = 2$) is given by the sum of the probabilities of configurations of the pair $(\hat{x}_2, \hat{x}_7)$ where $\hat{x}_2 = 0$, i.e. $(0, 0)$ and $(0, 1)$. Since the table describes the probabilities of the configurations, given the different congestion levels (and not the other way around), we do not have to do anything else than to sum up the probabilities in the cells where the column is $y = 2$, and the row is $\hat{x}_2 = 0$. That is: $p(\hat{x}_2 = 0|y = 2) = p(\hat{x}_2 = 0, \hat{x}_7 = 0|y = 2) + p(\hat{x}_2 = 0, \hat{x}_7 = 1|y = 2) = 0.81 + 0.03 = 0.84$

References

[1] A. Tsanas and A. Xifara, "Energy efficiency data set," 2012. data retrieved from UCI machine learning repository, `https://archive.ics.uci.edu/ml/datasets/Energy+efficiency`.

[2] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and Buildings*, vol. 49, pp. 560–567, 2012.

[3] T. Catalina, J. Virgone, and V. Iordache, "Study on the impact of the building form on the energy consumption," *12th Conference of International Building Performance Simulation Association, Sydney, 14-16 November.*, pp. 1726–1729, 2011.