ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

CESENA CAMPUS

DEPARTMENT OF ELECTRICAL, ELECTRONIC, AND INFORMATION ENGINEERING

"GUGLIELMO MARCONI"

SECOND CYCLE DEGREE IN BIOMEDICAL ENGINEERING

Class: LM-21

THESIS TITLE

INDOOR WELLNESS: A MACHINE LEARNING TECHNIQUE FOR INTERNET OF THINGS SENSOR DATA

Graduation thesis in

SENSORS AND NANOTECHNOLOGY

Supervisor
Prof. Marco Tartagni

Candidate
Gaetano Ingenito

Co-Supervisor
Prof. Aldo Romani
PhD Oumaima Afif

Academic Year 2022/2023

Contents

Introduction

The increasing levels of stress affects both the physical and mental well being of the people [1], this is mostly reflected in people working in closed spaces where almost 90% of employees have suffered from burnout syndrome in their lifetime.
As defined by the WHO, the quality of life is: "an individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns" [2]. Different aspects of someone's life can affect their well being, namely the social situation, physical and health status, financial and occupational condition and environmental surroundings. The latter one is also referred as environmental wellness and is influenced by the set of physical and chemical components of the environment.
It is a known fact that the psychological state of an individual affects its productivity and the associated risk of committing errors, in the biomedical field this has an higher weight due to the elevated chance of impacting in a negative way on the patient's health.
In this paper is analyzed how the environment affects the wellness of an individual and is shown how machine learning can help predict the emotional state of a person. The project involves the use of two Taunito sensors, courtesy of TAUA s.r.l., in different rooms of the University of Bologna - Cesena Campus. The devices collect the data which is analyzed through machine learning methods to assess the correlation between wellness and environmental parameters, to classify between comfortable rooms and non, to build a model able to predict the overall satisfaction level of the people in the room and to build a model able to predict the optimal temperature to keep in a room for which the satisfaction is maximized.

Chapter 1

PRIOR KNOWLEDGE

In this chapter are described the general concepts useful to better understand the scope of the project and the state of art technologies.

1.1 Wellness concept

Wellness is a broad concept used to describe a state beyond the absence of illness, but rather aims to optimize well-being [3]. In the 1970s a model was developed by Bill Hettler [4], a doctor at the University of Wisconsin, to describe this idea which included six factors, also known as dimensions, influencing the state of an individual. Since then more studies focused on the task, adding more dimensions and refining the original ones, nowadays eight different dimensions are recognized Fig.[1.1]:

- Emotional
- Spiritual
- Intellectual
- Physical
- Environmental
- Financial
- Occupational
- Social

Figure 1.1: Eight dimensions of wellness as defined by Bill Hettler, each contributes to the psycho-physical state of the individual. [5]

Wellness recognizes the interconnected nature of its various dimensions, understanding that an imbalance in one aspect can impact overall well-being. Achieving wellness is an evolving journey that requires intentional effort, self-awareness, and continuous personal growth. Numerous individuals, healthcare professionals, and organizations advocate for wellness, promoting it as a constructive method for leading a healthier and more satisfying life.
In the years the idea of wellness was spread in the corporate world and resulted in the creation of many employee assistance programs, in the US the clear example of aidance was the establishment in March 2010 of the ”Affordable Care Act (ACA)”, also known as ”Obamacare” whose key components are the imperative for all american citizens to have an health insurance coverage, expansion of Medicaid, a joint federal and state program that provides health coverage to low-income individuals and families, protection for pre-existing conditions in which insurance companies are prohibited to deny coverage, enstablisment of health insurance marketplaces, also known as exchanges, where individuals and families can purchase private health insurance plans allowing for a fast comparison between different plans.
Comfort as a basis for setting environmental standards in public buildings has developed out of recognizing that people need to be more than simply healthy and safe in the buildings they occupy. The idea of environmental comfort establishes a connection between the psychological preferences of workers in their environment and tangible outcomes, including enhanced

task performance. It also correlates with organizational productivity by ensuring that the workspace supports tasks related to work. The Vischer's habitability piramid Fig.[1.2], shows how enviromental comfort is comprehensive of at least three hierarchically related categories:

- Physical comfort: it includes the human's basic needs such as safety and hygiene. This is achieved through the use of standards and codes in architectural design and construction decision making
- Functional comfort: it refers to the ergonomic support for the user such as proper ventilation, lightning and ergonomic furnitures.
- Psychological comfort: it derives from a sense of belonging, ownership, and control over one's workspace

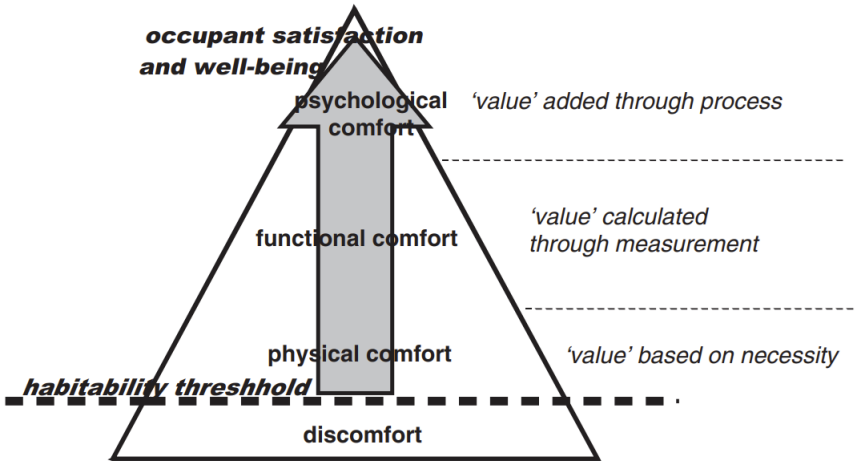


Figure 1.2: Habitability Piramid by Vischer showing the three hierarchical categories of comfort. The habitability threshold is the minimum requirement for an individual to inhabit an environment.

Based on the model, physical comfort is at the threshold of acceptable work-space, but actual well-being is achieved through the optimization of all the other categories. In the other cases, when one of the aspects is lacking, stress start to accumulate and has a significant and wide-range effect on workers, especially those working in high-pressure and demanding fields as healthcare. Stress is a natural and adaptive reaction that can be triggered by various situations or events, often referred to as stressors which can be physical, emotional, psychological, or environmental in nature. Stress is a complex response involving both physiological and psychological components whose main effects, sorted for importance, are:

1. Physical health issues (exhaustion and weakened immune system)

2. Mental health issues (burnout and depression)
3. Higher risk of substance abuse
4. Impact on the work's target
5. Work-Life balance challenge

Stress is the result of a complex system affected by multiple factors, in Fig.[1.3] is shown a conceptual framework describing how the physical environment may set in motion a process leading to stress.

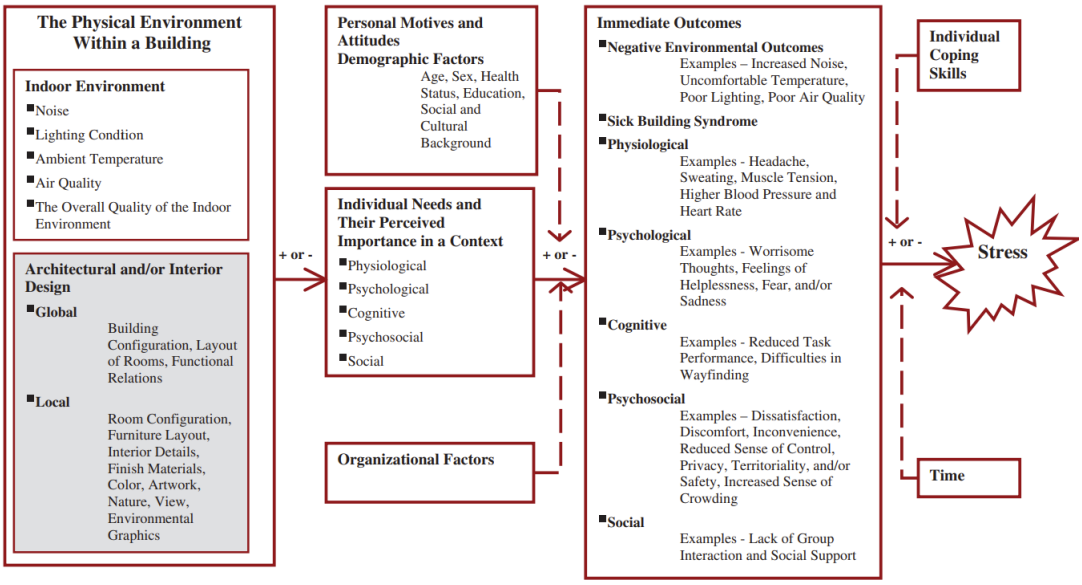
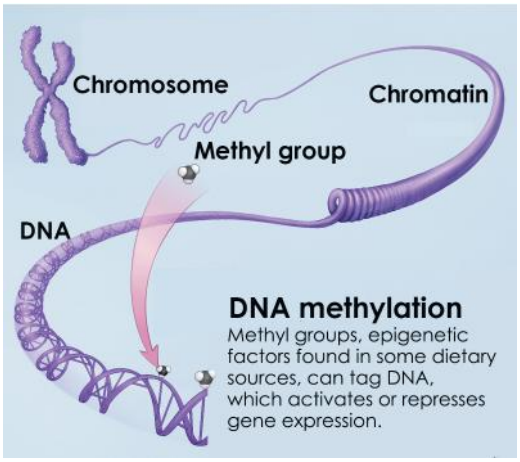


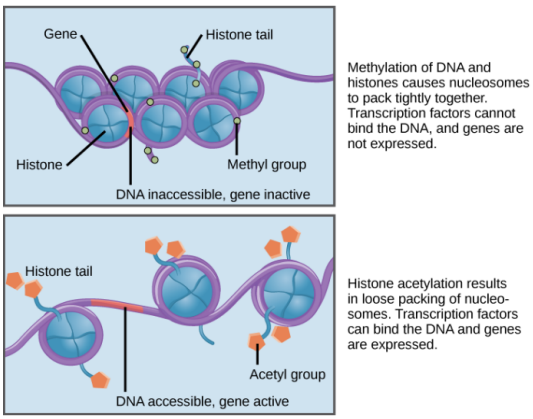
Figure 1.3: Framework describing the various processes leading to stress [6]

1.2 Epigenetics

The influence of light exposure, weather and air pollution on a person's mood is widely recognized, as an example people living in rainy regions are more likely to suffer from seasonal depression than those living in mild climate zones and people living in polluted cities have an higher risk of developing a tumor. Those elements are example of epigenetics factors.
The difference between the genotype and phenotype was discovered in 1911 by Wilhelm Johannsen, although the meanings have evolved since they were first introduced. Genotype refers to the set of genetic information encoded in the DNA strands while phenotype also includes the effect of the surroundings on the gene expression and the various mechanisms influencing it.
In biology, epigenetics (from the greek "on top of origin") are the stable heritable traits that cannot be explained by the change in DNA sequence and usually are related to changes in the gene expression due to environmental factors, diet and drugs usage.
The main mechanisms affecting the gene expression are DNA methylation and histone modification. The first acts by adding a methyl group ($-CH_3$) to either adenine or cytosine resulting in a repression of the coding gene Fig.[1.4a], while the second one acts by attaching a molecule to an histone tail altering the extent to which the DNA is wrapped around and, consequently, the availability of genes. In particular when a methyl group is added, the histones get closer to each other and the gene expression is reduced, on the other hand adding an acetyl group ($-COCH_3$) enhances the expression Fig.[1.4b].



(a) DNA Methylation mechanism, the $-CH_3$ binds to cytosine or adenine and modifies the availability of the gene [7]



(b) Histone modification, the genetic repression (enhancing) is mediated by $-CH_3$ ($-COCH_3$) which binds to the histone tails and modifies the distance [8]

Figure 1.4: Epigenetic mechanism acting on the gene expression

Stress is an indirect epigenetic factor [9],[10] which acts on the organism through cortisol, a steroid hormone associated with mood disorders [11]. Cortisol secretion is regulated by the hypothalamic hormone, CRH, and the pituitary hormone, ACTH, in the hypothalamus-pituitary-adrenal axis, its deficency induces Addison's disease, while overproduction is related to Cushing's syndrome. The main role of cortisol is to set the body in a fight-flight response by acting on the metabolism of both fat and sugar, regulating blood pressure and enhancing the immune system.
Other types of epigenetic factors are:

- Light exposure: the amount of light affects the circadian rhythm which acts through the suprachiasmatic nucleus (SCN) in the hypothalamus, in particular suppresses the production of melatonin and increases the levels of cortisol. Disruptions in this cycle, as seen in conditions like shift work or irregular light exposure, can lead to metabolic imbalances and health issues such as obesity, insulin resistance and metabolic syndrome.

- CO_2 concentration: it plays a vital role in the acid-base balance, oxygen transport and respiratory functions, both physiological effect interconnected overall with the wellness of the subject. Studies show how an excessive amount of CO_2 decreases the cognitive functions [12],[13].

- Volatile Organic Compounds: also known as VOC, they are molecules mostly produced through industrial and biological processes. Most of them do not have acute toxicity, but have a long-term chronic health effect and represent a potential health and environmental threat [14]. VOC are not directly responsible of epigenetic changes, but can influence them through various pathways.

- Temperature and Humidity: high levels of temperature and humidity can induce irritability and discomfort, on the other hand low levels induce feelings of sadness and depression [15]. Some species alters their gene expression patterns to enhance survival and cope with temperature fluctuations.

1.3 IoT

The term ”Internet of Things” (IoT) refers to devices equipped with sensors, processing capabilities, software, and other technologies that establish connections and exchange data with other devices and systems through the Internet or other communication networks [16]. The evolution of this field is a result of the convergence of various technologies, such as ubiquitous computing, commodity sensors, powerful embedded systems and machine learning. Older disciplines like embedded systems, wireless sensor networks, control systems, and automation (including home and building automation) independently and collectively contribute to the realization of the Internet of Things.
In the consumer market IoT technology is commonly associated with ”smart home” products including devices and appliances like lighting fixtures, thermostats, home security systems, cameras, and other household items that support one or more standard ecosystems which can be controlled through devices linked to that ecosystem, such as smartphones and smart speakers. Additionally, IoT plays a significant role in healthcare systems.
The Internet of Medical Things (IoMT) is an application of the IoT for medical and health-related purposes, data collection and analysis for research and monitoring. These monitoring tools encompass a wide range, from basic blood pressure and heart rate monitors to more sophisticated devices capable of piloting specialized implants like pacemakers, Fitbit electronic wristbands, or advanced hearing aids. Some hospitals have implemented ”smart beds” capable of autonomously detecting occupancy and patient movements to rise. These beds can adjust themselves to provide optimal pressure and support, eliminating the need for manual intervention by nurses.
The concept of the Internet of Things (IoT) was initially introduced in 1982 when a modified Coca-Cola vending machine was able to report its inventory and the temperature status of newly stocked drinks. The term ”Internet of Things” was coined by Kevin Ashton of Procter & Gamble in 1999, since then the proliferation of IoT devices has grown exponentially, as depicted in Fig.[1.5].

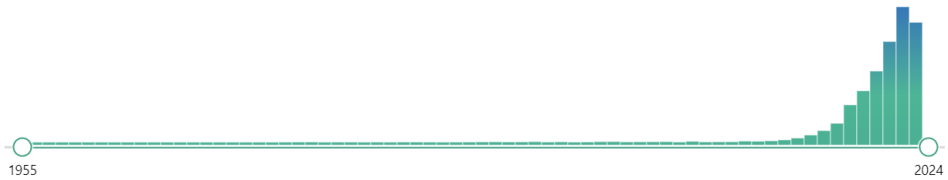


Figure 1.5: Growth trend of Internet of Things devices

Despite the name, the Internet of Things is somewhat of a misnomer, as devices don’t

necessarily have to connect to the public internet, but merely need to be linked to a network and possess individual addressing. The main used communication protocols [17]-[18] are:

- AMQP (Advanced Message Queuing Protocol) serves as an open standard protocol designed for facilitating message-oriented middleware. Notable elements of AMQP encompass message queuing, routing, reliability, and interoperability. It establishes a framework of regulations and practices governing the production, consumption, and exchange of messages, thereby enabling smooth communication between diverse applications.

- Bluetooth is a short-range wireless technology that uses short-wavelength, ultrahigh-frequency radio ($2.4GHz$) waves mostly used for audio streaming, but is has become a significant enabler of wireless and connected devices. It follows a master-slave architecture, whereas one device (master) manages the connection with up to seven other devices (slaves). The key features of Bluetooth include low power consumption, simplicity of use and support for various applications such as audio streaming, data transfer, and device control.

- Cellular is one of the most widely available and well-known options available for IoT applications, and it is one of the best options for deployments for long range communications. It's capable of sending high quantities of data, however those features come at a price: higher cost and power consumption than other options.

- CoAP for Constrained Application Protocol is designed to work with HTTP-based IoT systems. It relies on User Datagram Protocol to establish secure communications and enable data transmission between multiple points even in conditions of low bandwidth and low energy devices.

- DDS for Data Distribution Service is a middleware protocol and API standard for data-centric connectivity which integrates the components of a system together providing low-latency, data connectivity, extreme reliability and a scalable architecture making it suitable for real time performances in distributed systems.

- LoRa is a non-cellular long range wireless communication protocol. Differently from the cellular it covers a longer distance and has a lower power consumption at the cost of a lower data rate.

- LWM2M for Lightweight M2M (machine to machine) is specifically designed for remote device management and telemetry being crafted to be lightweight, efficient and

suitable for resource-constrained devices. It follows a client-server architecture where IoT devices act as clients and connect to LWM2M servers enabling an efficient communication and management of devices in a scalable manner.

- MQTT for Message Queuing Telemetry Transport is designed to be a low bandwidth, high latency on unreliable networks. Its simple messaging protocol works with constrained devices and enables communication between multiple devices. MQTT provides a "Last Will and Testament" option, allowing users to specify a message that the broker should publish on their behalf if the client unexpectedly disconnects. This feature is useful for detecting and responding to device failures.

- Wi-Fi is designed for short to medium range distances, particularly suited within LAN environments. It operates in the $2.4GHz$ and $5GHz$ frequency bands.

- XMPP for Extensible Messaging and Presence Protocol is based on Extensible Markup Language (XML) and used for real time exchange of structured but extensible data between multiple entities on a network.

- Zigbee is a mesh network protocol designed for low power, low data rate and short range communication, making it suitable for building and home automation applications. Zigbee uses a mesh networking topology, allowing devices to communicate with each other through a network of nodes. This improves reliability and extends the effective range by allowing messages to hop through the intermediate nodes.

- Z-Wave is a mesh network protocol designed for home automation and control applications. Differently from Zigbee it operates in the sub-1Ghz frequency range offering better wall penetration, but with a lower data rate.

No single IoT communication protocol is best, nor is any one right for every deployement.
In the following section is described the Taunito sensor by TAUA s.r.l. used in the project.

1.3.1 Taunito sensor

The Taunito INAIR sensor is a patented device by Taua s.r.l., it is a IoT device designed to measure various environmental parameters in a closed space. The design is a oval shaped device which comes in different colors, it does not require any kind of installation procedure and thus no buttons. It can be placed either indoor or outdoor, in moving vehicles or in still places thanks to the long duration 5V battery.

Figure 1.6: Taunito INAIR device placed in the control room of the ”Alma Mater Studiorum - Università di Bologna - Cesena Campus”

The inner part of the device comprehends an electronic card onto which are attached the sensors whose data is analyzed, filtered and sent to the online server via e-sim. The measured quantities are described in the table below [1.1].

Detection Time	Battery level
Humidity	Temperature
Pressure	CO_2 Concentration
GPS coordinates	Indoor Air Quality (IAQ)
Static IAQ	Breath VOC Concentration
Gas Percentage	Accuracy

Table 1.1: Physical quantities measured by the various sensors of the TAUNITO device

The quantities are gathered through two principal sensors:

- BME680 by Bosh, it is a 4-in-1 sensor with gas, humidity, pressure and temperature measurement. The sensor module is housed in an extremely compact metal-lid LGA package of only $3.0 \times 3.0\ mm^2$ with a maximum height of $1.00\ mm$. Each module can be independently enabled or disabled, the operating range spans from $-40°C$ up to $+85°C$, 0% up to 100% of relative humidity and $300hPa$ up to $1100hPa$.
  A software solution (BSEC: Bosch Software Environmental Cluster) provides an index for air quality (IAQ) which ranges in the interval $[0-500]$ with a resolution of 1. Based on its values, the German Federal Environental Agency classifies the ambient in seven classes Fig.[1.7].

IAQ Index	Air Quality	Impact (long-term exposure)	Suggested action
0 – 50	Excellent	Pure air; best for well-being	No measures needed
51 – 100	Good	No irritation or impact on well-being	No measures needed
101 – 150	Lightly polluted	Reduction of well-being possible	Ventilation suggested
151 – 200	Moderately polluted	More significant irritation possible	Increase ventilation with clean air
201 – 250[9]	Heavily polluted	Exposition might lead to effects like headache depending on type of VOCs	optimize ventilation
251 – 350	Severely polluted	More severe health issue possible if harmful VOC present	Contamination should be identified if level is reached even w/o presence of people; maximize ventilation & reduce attendance
> 351	Extremely polluted	Headaches, additional neurotoxic effects possible	Contamination needs to be identified; avoid presence in room and maximize ventilation

Figure 1.7: Classification designed by the German Federal Environental Agency: high scores of IAQ correspond to polluted enviroments, on the other hand low scores correspond to healthy ones.

The algorithm returns both the ”IAQ” and ”ST-IAQ” index, the first is the simple metric mentioned above, the second is a filtered (trimmed) version of the first best representative of the measurement in case the device is fixed in place.
BME680, being a metal oxide-based sensor is able to detect VOC by asdorption (and

subsequent oxidation/reduction) on its sensitive layers thus making it capable of measuring the sum of VOC in the local area. The higher the concentration of reducing VOC, the lower the resistance and viceversa. Unfortunately the sensor is not able to distinguish between the different types of VOC such as benzene, formaldehyde, toluene or xylene, but only offers a cumulative measurement of those.

- ExplorIR-M by GSS is a robust, low power CO_2 sensor in miniature format. It uses nondispersive infrared lamp to direct waves of light through a tube filled with a sample of air. The air moves toward an optical filter in front of an IR light detector which measures the amount of IR light that passes through the optical filter. In an NDIR CO_2 sensor, the band of IR radiation produced by the lamp is very close to the 4.26-micron absorption band of CO_2 which serves as a signature or "fingerprint" to identify the molecule. As the IR light passes through the sample tube of air the gas molecules absorb the specific band of IR light while letting other wavelengths of light pass through. At the detector end, the remaining light hits an optical filter that absorbs every surviving wavelength of light and is read by an IR detector. Comparing the original spectrum and the absorbed one is possible to evaluate the amount of CO_2 in the sample.

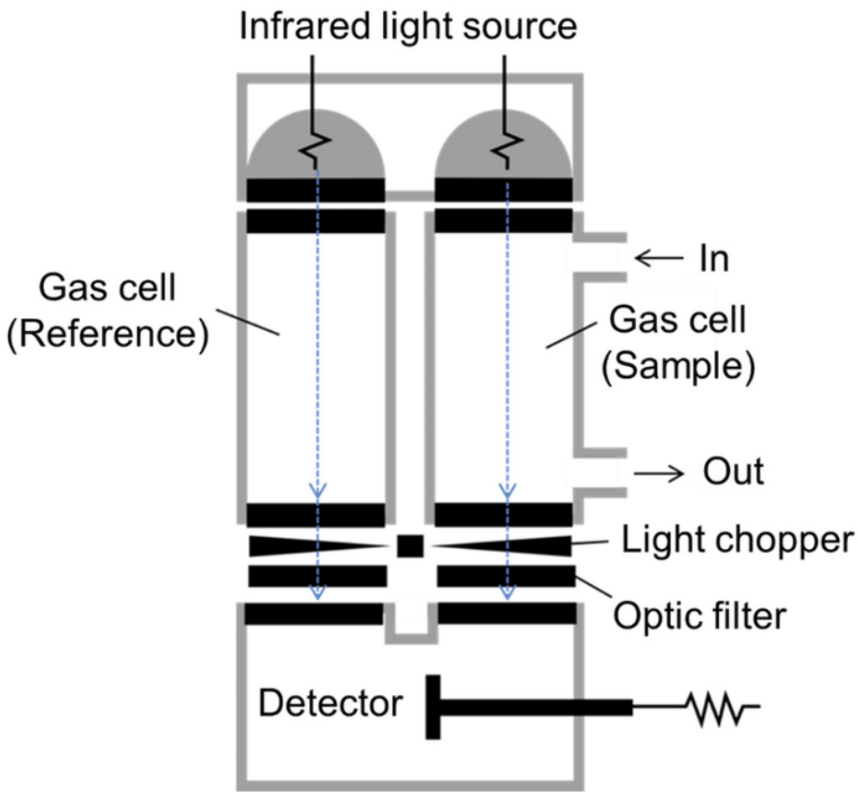


Figure 1.8: Schematic working principles of infrared molecules detection [19]. The measured spectrum at the output is compared to the reference one to predict the CO_2 concentration

1.4 State of art

The interest in the wellness of the individual based on the environmental factors is a growing, rather new approach in the biomedical field. Up until the last decade research has been more focused on the idea of solving a problem when already there rather than trying to find ways to prevent it from happening.

One example of preventive research is the study conducted by Bauxell et al. [20] in which, through a machine learning approach they were able to predict the risk of dengue fever, a potentially fatal pathology carried by mosquitoes. This was done by combining both environmental data and socioeconomical data.

The modern research has been more focused on children's wellbeing in school environment as shown in [21],[22] and [23], both on the health and psychological side. Those works show how various environmental parameters, mainly CO_2 and VOC concentration, are associated with fluctuations in cognitive functions. One great limitation of those works is the use of data gathered from different sources which introduces errors both for the different calibration and accuracy and for the filtering techniques, nevertheless the studies confirm how a suitable environment helps increasing the comfort and productivity of the occupants.

A study conducted on the same population [24], but centered on the medical conditions showed how the environment affects the health status where polluted air increases the risk of contracting pathologies.

The combination of biosensor data and environmental condition aimed at the psychological wellness was the main ingredient of the study conducted by Van Der Valk et al. [25], in their work the goal was to check for a correlation between the data and the productivity of the employees. The results showed a promising relationship, but the ensemble of participants was very strict and moreover for the study to be as accurate as possible the attendees had to stay sit in place, possibly compromising the scores. An analogous objective is the one conducted by Ma et al. [26] in which a mobile app was developed to keep track of the phone's owner mood based on various factors including light exposure, allowing mobility of the user. Many of the used features were extracted by mobile and communication data such as acceleration, text and call frequency, sound intensity and location. The model built followed a Naive-Bayes classification algorithm and proved an accuracy close to 50%.

The framework presented in this paper differs from the above mentioned since it is based on physical and psychological variables gathered from a bigger sample population and aims at optimizing the well-being using simpler techniques.

Chapter 2

METHODS

In this chapter are described both the mathematical tools and the psychological test used during the analysis. The first falls back in the machine learning scope, while the second is a way to quantify the level of satisfaction.

2.1 Machine learning approaches

The goal is to identify, if it exists, the relationship between the physical environment and the well-being of the subjects through machine learning. The methods used are ”Principal Component Analysis” (PCA), ”Soft Independent Modelling of Class Analogies” (SIMCA), ”Projection to Latent Structures” (PLS) and ”Local Weighted Regression” (LWR). PCA is used as a dimensionality reduction method to identify the main features which best influence the satisfaction scores and to look for a statistical difference between the class groups, SIMCA is used to classify observations based on the features extracted via PCA, PLS is used to build a regression model able to predict a (latent) variable of the system and lastly LWR is used to optimize the regression since some non linearity is expected to exist. In the following sections are also showed ”Multiple Linear Regression” and ”Principal Component Regression”, two regression methods that help to better understand the PLS one.
A latent variable is a variable that can only be inferred indirectly through a mathematical model from other observable variables that can be directly observed or measured [27], some examples are the overall health status of a patient or its happiness which are a combination of different aspects. In the present work the latent variable is the satisfaction level.
All the methods mentioned above require one (for PCA and SIMCA) or two (for PLS ans LWR) datasets, the environmental data is stored in a $[N \times K]$ matrix called X, where N is the number of independent observations, K is the number of the variables used to describe the observations. The satisfaction metric is stored in a $N \times 1$ matrix called Y.

The two dataset are graphically represented in Fig.[2.1].

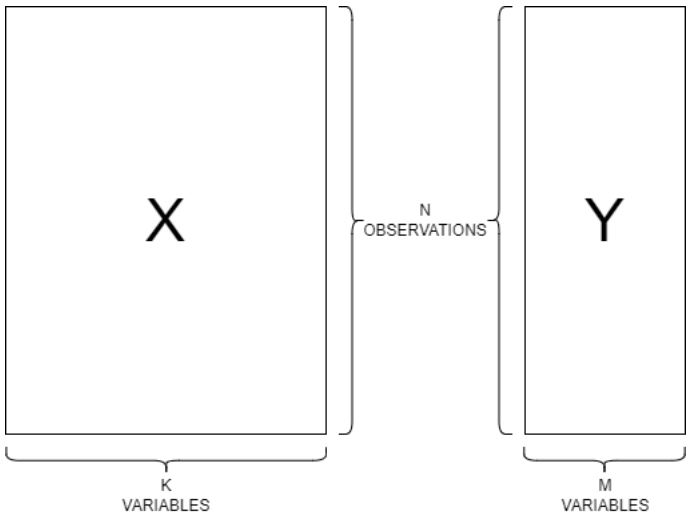
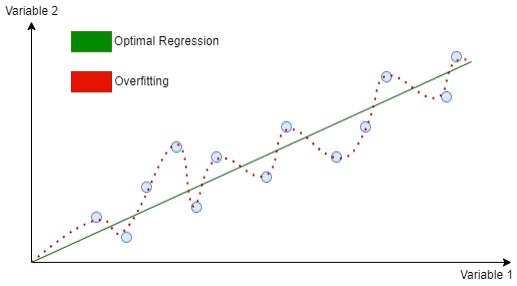


Figure 2.1: Illustration of the input data matrix X and the output vector Y

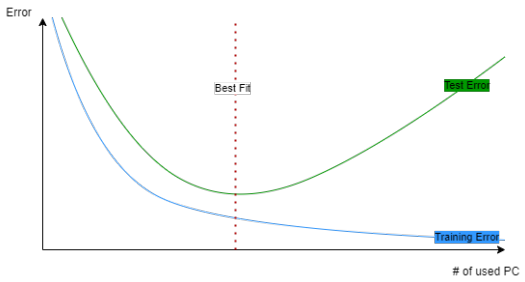
The above mentioned methods are based on the hypothesis of linearity between the input (X) and output (Y) of a model constructed on the training set. It is important to remember that a model is always an approximation of the real system from where the data comes from and its reliability can be evaluated with different statistical coefficients [2.1], mainly:

$$\begin{aligned} \text{Root Mean Square Error:} \quad & RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \\ \text{Coefficient of determination:} \quad & R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \\ \text{Squared Prediction Error:} \quad & SPE = \sum_i (y_i - \hat{y}_i)^2 \end{aligned} \tag{2.1}$$

Where $\hat{y}$ is the estimated output of the model and $\bar{y}$ is the mean value of the observed data.
In particular, the R^2 coefficient describes how well the model performs in explaining the variance of the dataset. When close to 1 it means that the model's regressors are able to well explain the data.
One of the main problems when working with machine learning algorithms is "overfitting" in which the model turns selective with respect to the dataset and is not able to handle generality Fig.[2.2a]. This effect is faced when the model complexity, meaning the number of principal components used, is too big. One way to check for it is to look at the test and training error Fig.[2.2b] as a function of the model complexity.



(a) Difference between the optimal regression curve (green) and the overfitted one (red). When overfitting occurs the model becomes too dependant on the training data and loses the ability to generalize



(b) Effect of the overfitting on the error, increasing the number of PC used to represent the model the training error (blue) decreases, but the test error (green) increases

Figure 2.2: Visual representation of overfitting

The choice of the best number of regressors is done via cross-validation, in which the entire dataset is split into two: a train and test dataset. The rule of thumbs splits the dataset in a $80 - 20\%$ with respect to the original data. The train test is used to calibrate the model and is then applied onto the test set. In both cases the error is computed and compared.
The machine learning algorithm used are described in the following sections.

2.1.1 Principal Component Analysis (PCA)

PCA is a common statistical technique used to reduce the dimensionality of the data keeping most of the information. The key idea is to transform the original reference system built on the K variables in a new one using only A linear combinations of the features called principal components which are chosen via cross-validation. Each observation is represented as a single point in the variable space, the new reference system is found by minimizing the distance between the points and the new axis, or equivalently, by maximizing the distribution of their projections.

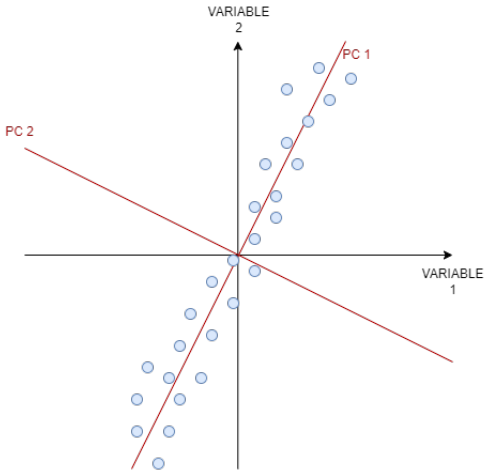


Figure 2.3: Change of the reference system, the basis of the new one aim at maximizing the variance of the projections of the points on the axis

The values of the projections, also known as scores, are stored in a $[N \times A]$ matrix called T while the components of the new axis in the original reference system, also knows as loadings, are stored in a $[K \times A]$ matrix called P. The relation [2.2] holds between the two.

$$T = XP \tag{2.2}$$

Recall that the loading vectors are oriented in a way that the variance of the scores in the corresponding direction is maximal and the vectors are orthogonal, mathematically can be translated as in [2.3]:

$$\begin{aligned} \max \sigma^2 &= t_1^T t_1 = p_1^T X^T X p_1 \\ \text{s.t.} \quad & p_1^T p_1 = 1 \end{aligned} \tag{2.3}$$

The classical approach to compute the principal components is to apply the Lagrange multi-

plier method [2.4] since both a minimization problem and a constrain are present. For ease it is shown just for the first principal component, but it can be proven to work with the others.

$$\begin{aligned} \max \sigma^2 = 0 &= p_1^T X^T X p_1 - \lambda_1 (p_1^T p_1 - 1) \\ 0 &= 2X^T X p_1 - 2\lambda_1 p_1 \\ 0 &= (X^T X - \lambda_1 I_{K\times K}) p_1 \\ X^T X p_1 &= \lambda_1 p_1 \end{aligned} \tag{2.4}$$

The last equation shows that the principal components and the loadings are respectively the eigenvalues and eigenvectors of the correlation $X^T X$ matrix.

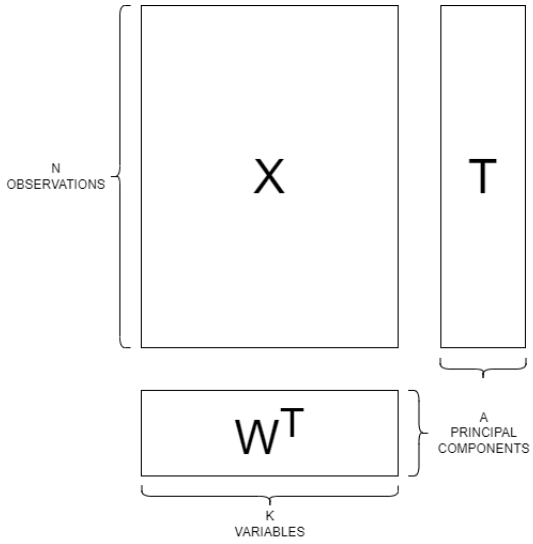


Figure 2.4: Illustration of the input data X, score matrix T and loading matrix W

The proportion of variance that each eigenvector explains can be computed dividing the corresponding eigenvalue by the sum of all the other ones, for this reason only the biggest terms are kept which are able to explain most of the data. Increasing the number of PCs also increases the explained variance, but an excessive number models noise too leading to overfitting. When present, the training error decreases while the test error increases.
To avoid the problem a class of iterative model validation techniques, also known as cross validation is used. Generally for PCA the preferred one is "Wold CV" [28] which consists in two loops, the outer one iterates along the number of principal components $a = 1, ..., min\{N, K\}$ and the inner one divides the dataset in G groups. At each run all the dataset expect for the $g - th$ group is included in a matrix called X_{-g} and a PCA model is built on it. With the results of the analysis the $g - th$ group is predicted and the error $E_{a,g} = |X_g - \hat{X}_{a,g}|$ is computed. This process is iterated for all the G groups and at the end the residual matrix E_a is

built containing all the errors, the Q_a^2 coefficient is computed, mathematically equal to R^2, but with two main differences:

- The Q^2 coefficient is always less than the R^2 due to the fact that the first is computed on a smaller version of the dataset, thus hiding from the model some variability
- The Q^2 coefficient decreases after a certain number of PCs is used while the R^2 is always increasing, this indicates that the latest added one is not systematic

This process is iterated along the number of PCs and the best score, corresponding to the best number of components, is chosen. A pseudocode of the process is shown below.

Algorithm 1: Wold Cross-Validation method (pseudocode)

for $a = 1 : min\{N, K\}$ **do**
  for $g = 1 : G$ **do**
    X_{-g} = splice(X,g_{th} group) ;
    [$\sim$,P] = PCA(X_{-g},a);
    $T_g = X_g * P$;
    $\hat{X}_{a,g} = T_g * P^T$;
    $E_{a,g} = abs(X_g - \hat{X}_{a,g})$
  end
  $E_a = [E_{a,1}, E_{a,2}, ..., E_{a,G}]$;
  $Q_a = 1 - \frac{\sigma^2(E_a)}{\sigma^2(X)}$;
end

One advantage of using PCA is that it provides the Hotelling's T^2 coefficient [2.5] able to show the reliability of the prediction.

$$T^2 = \sum_{a=1}^{A} \left(\frac{t_{i,a}}{\sigma_a} \right)^2 \tag{2.5}$$

Where σ_a is the standard deviation of each component. Geometrically the T^2 coefficient is the distance from the center of the (hyper)plane to the projection of the observation onto the (hyper)plane.
The classical eigenvalue approach is frequently not used due to the high computational cost and due to the fact that it computes all the eigenvalues even if only the first A are needed. Two alternate algorithms are described:

- **Singular Value Decomposition**

The SVD algorithm factorizes the data matrix as in [2.6]

$$X = U\Sigma V^T \tag{2.6}$$

where U and V are orthonormal matrices, respectively $[N \times N]$ and $[K \times K]$, and Σ is the $[N \times K]$ diagonal matrix containing the square root of the eigenvalues of the $X^T X$ matrix, also known as singular values.
The relationship to principal component analysis is shown in [2.7]:

$$\begin{gathered} X = TP^T \\ T = U\Sigma \quad \text{and} \quad P = V \end{gathered} \tag{2.7}$$

The columns U and V are the eigenvectors of respectively XX^T and $X^T X$.
The main advantages of using SVD rather than the traditional eigenvalues method are the assured numerical stability and the fact that the principal components are already sorted by importance.

- **Non-linear iterative partial least squares**
  The NIPALS algorithm is a sequential method that allows to compute the principal components. It allows to terminate the computation after a fixed number of PCs have been found and is also able to handle missing data, for these reasons it is one of the most used algorithms in most of the software packages.
  The steps of the algorithm are:

  1. Randomly initialize a column t_a
  2. For each column of X a LS regression is performed onto the vector t_a as in [2.8]

     $$p_a = \frac{t_a^T X_k}{t_a^T t_a} \tag{2.8}$$

     The regression coefficient is stored in a vector p_a
  3. Normalize the magnitude of the vector p_a
  4. For each row of X a LS regression is performed onto the vector p_a as in []

     $$t_a = \frac{X_i p_a}{p_a^T p_a} \tag{2.9}$$

     The regression coefficient is stored in a vector t_a
  5. Iterate the steps 2,3,4 until convergence

6. Deflate the X matrix, meaning that the variability of the a component is removed. This is done as in [2.10]

$$E_a = X_a - t_a p_a^T$$
$$X_{a+1} = E_a \tag{2.10}$$

7. Increment a and start again from step 2 until the wanted number of PCs is obtained

Below in figure [2.5] is shown the graphical representation of the NIPALS algorithm for PCA.

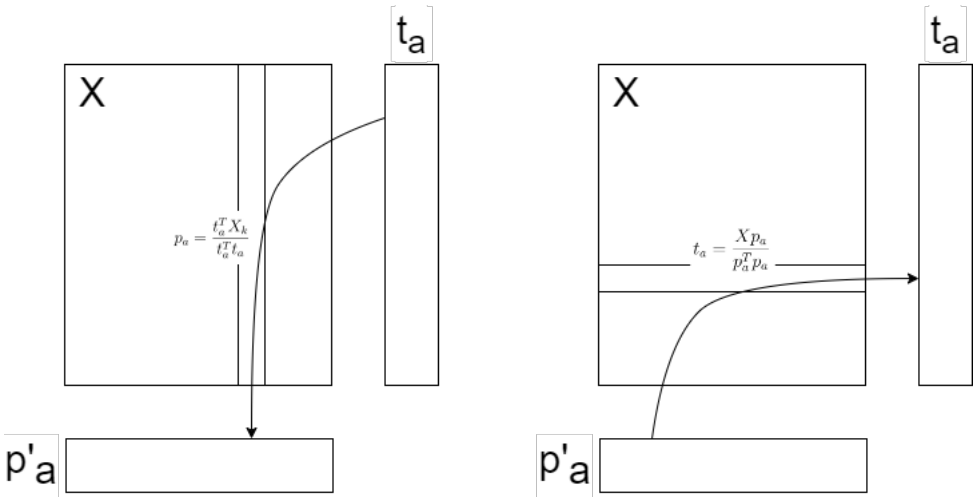


Figure 2.5: Graphical representation of the NIPALS algorithm, each arrow corresponds to a least square regression

2.1.2 Soft Independent Modelling of Class Analogies (SIMCA)

Soft Indipendent Modelling of Class Analogies is a statistical method for supervised classification of data. For a given class, all the samples belonging to it are analyzed through PCA and only the most representative components are retained. For each sample in the modelled class, the Hotelling's T^2 coefficients are measured and its distribution is built. The critical distance represents a threshold beyond which a sample is considered an outlier to the class and is often defined as a certain percentile of the Hotelling's T^2 distribution, generally the 95th or 99th percentile. The Hotelling's T^2 distribution [2.11] is defined as:

$$T^2 = (\bar{x} - \mu)^T W^{-1} (\bar{x} - \mu) \tag{2.11}$$

where $\bar{x}$ is the sample mean, μ is the population mean and W is the sample covariance matrix [2.12] defined as:

$$W = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \tag{2.12}$$

To classify new samples, the Hotelling's T^2 scores are computed and compared against the critical distances of each class distribution, if it does not match with any of them then it is considered to be an outlier and unclassified.
The main tool to look at to assess the efficency of the algorithm is the confusion matrix, a $[N \times N]$ matrix, where N is the number of possible classes. Once the SIMCA algorithm is run a set of predicted class is returned, those are compared with the actual real classes to give an estimate of the accuracy of the algorithm. Each row of the matrix corresponds to the actual class, each column to the predicted one. The (i, j) entry of the matrix is computed as in [2.13]:

$$C(i,j) = sum(RealClass == i \;\&\; PredClass == j)/sum(RealClass == i) \tag{2.13}$$

The entries on the main diagonal represent the fraction of classes that are correctly identified, ideally are all equal to 1, on the other hand the other entries correspond to the case in which the real class i is wrongly identified as predicted class j.
Additional metrics to assess the accuracy in binary classification problems are the "Matthew correlation coefficient (MCC)", whose interpretation is similar to the Pearson's one, the P Precision coefficient and F_1 score. The MCC measures the association between two binary variables, given the confusion matrix two binary variables are positively associated if most of the data falls along the diagonal cells and negatively vice-versa. The MCC is computed as in

[2.14].

$$MCC = \frac{cs - \vec{t} \cdot \vec{p}}{\sqrt{s^2 - \vec{p} \cdot \vec{p}} \sqrt{s^2 - \vec{t} \cdot \vec{t}}} \tag{2.14}$$

where c is the total number of samples correctly predicted, s is the total number of samples, t_k is the number of samples belonging to class k and p_k is the number of times class k was predicted. Similarly to the Pearson's coefficient the ideal values is the one close to 1.
The Precision coefficient [2.15] is defined as the ratio between the number of true positive and the total number of positive predictions made by the model, it measures the percentage of truly positive out of all the positive predicted.

$$P = \frac{TP}{TP + FP} \tag{2.15}$$

The F_1 score [2.16] is the the harmonic mean of precision and recall (same as true positive ratio), taking account of both false positive and false negatives it performs well on imbalanced dataset.

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \tag{2.16}$$

2.1.3 Multiple Linear Regression (MLR)

The multiple linear regression algorithm is a tool able to predict the values in the Y matrix through the data stored in the X matrix.
The MLR algorithm is the easiest to use and to implement, it consists in finding the direct correlation between the data matrix X and the data matrix Y via a linear transformation mediated by the regression coefficients β shown in equation [2.17]. If Y has more than one column a different model has to be built for each of them.

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_N X_{iN} = \beta_0 + \sum_{j=1}^{N} \beta_j X_{ij}$$

In matrix form:

$$\hat{Y} = \beta_0 + X\beta \tag{2.17}$$

The regression coefficients β are those that minimize the error between the output of the model $\hat{Y}$ and the real values Y, can be thought as the partial derivatives of the output of the model with respect to the corresponding variable. The β_0 coefficient is the intercept of the line in the feature space and can be omitted if, during the pre-processing step, the data is mean centered. The best fit is performed using the least-square method [2.18]

$$\hat{\beta} = \underset{\beta}{arg\,min}\ ||Y - X\beta||^2 \Longrightarrow \hat{\beta} = (X^T X)^{-1} X^T Y$$

Proof:

$$\Phi = ||Y - X\beta||^2 = Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta$$
$$\nabla\Phi = -2X^T Y + 2X^T X\beta = 0$$
$$2X^T X\beta = 2X^T Y$$
$$\beta = (X^T X)^{-1} X^T Y \tag{2.18}$$

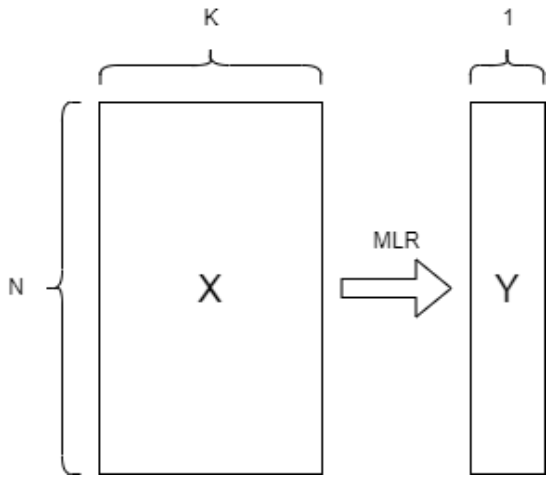


Figure 2.6: Graphical representation of MLR, the relation is mediated through the regression coefficients β

When using MLR the hypothesis to keep in mind are:

- A linear relationship between the X variables and the Y vector exists
- X columns are independent
- Data is noise free

The main limitations of using it is that it cannot be used if more than one variable has to be predicted, since it is required to build a model for each. In most of the data sets the variables are generally correlated, applying MLR can potentially lead to an ill-conditioned problem where a small variations in the X data, such as noise, leads to a big variation in the regression coefficient.

2.1.4 Principal Component Regression (PCR)

The idea of PCR is to overcome the limitations of simple MLR by applying it onto an orthogonal matrix able to summarize the original data matrix. The $[N \times A]$ matrix T used is an approximation of the original one, represented by just a fixed number A of principal components chosen via cross-validation and representative of the variation of the data. Geometrically PCA is a roto-translation of the reference system whose goal is to minimize the distance between the observations and the new axis, or equivalently, to maximize the distribution of the projections. The values of the projections, also known as scores, are stored in the T matrix while the components of the new axis in the original reference system, also knows as loadings, are stored in a $[K \times A]$ matrix called P. The relation between the two is described in [2.19].

$$T = XP \tag{2.19}$$

The MLR algorithm is then applied onto the T matrix in order to predict the values of Y [2.20]

$$\hat{Y} = T\beta \tag{2.20}$$

Where β are the regression coefficients that minimize the error between the model and the data computed via least-square.

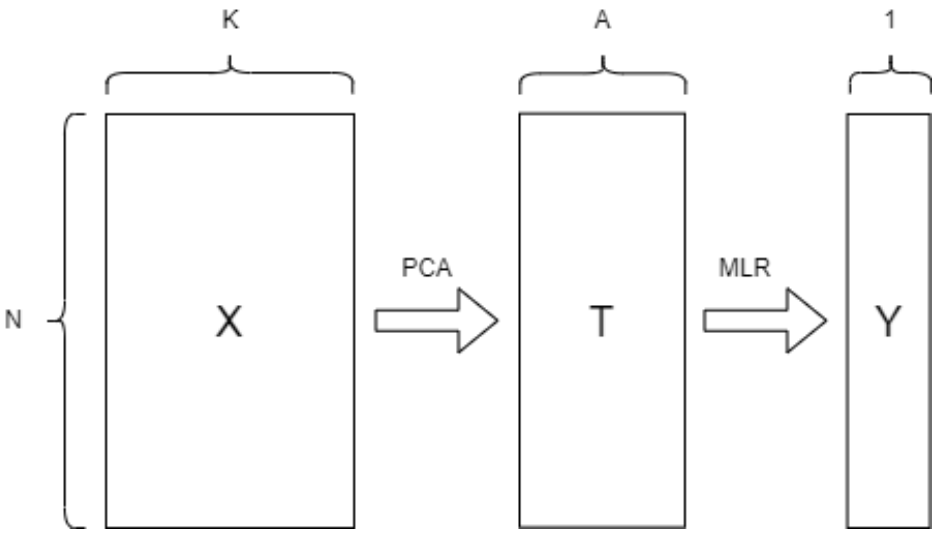


Figure 2.7: Graphical representation of PCR, the relation is first mediated through PCA and then through MLR

Using the PCR algorithm rather than the MLR one the orthogonality of the T matrix is assured, thus making it always invertible and it also bypasses the noise problem. The main

limitation is that for each column of Y a different PCR model has to be built.

2.1.5 Projection to Latent Structures (PLS)

PLS is a common statistical technique used to build a model able to predict a latent variable of the system in analysis. The PLS method aims at maximizing both the variance of scores for X and Y, and simultaneously the correlation between the two.
Two sets of scores (T and U) and loadings (W and C) are computed [2.21]:

$$\begin{aligned} t_a &= X_a w_a \quad \text{for the } X\text{-space} \\ u_a &= Y_a c_a \quad \text{for the } Y\text{-space} \end{aligned} \tag{2.21}$$

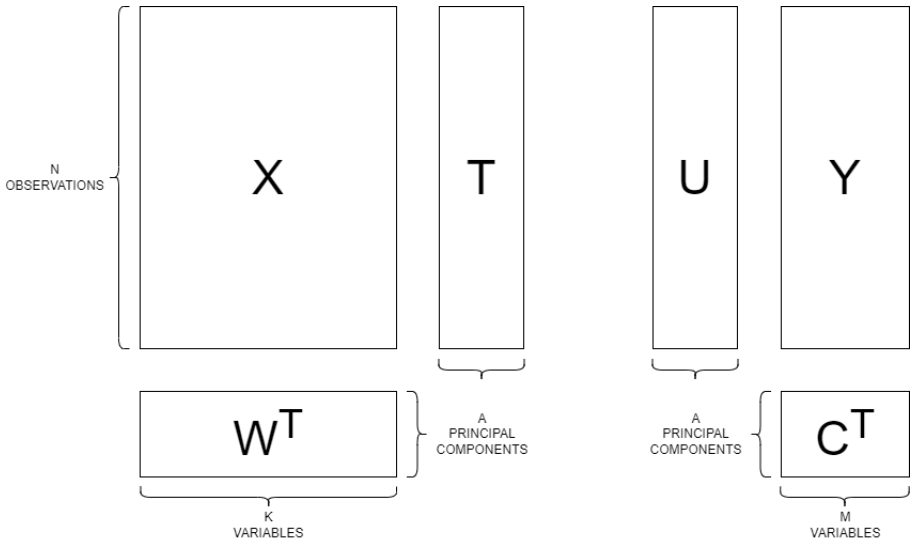


Figure 2.8: Illustration of the two datasets X and Y, the respective score matrices T and U and loading matrices W and C

The loadings are those for which the covariance [2.22] between the scores is maximized, it is shown how it depends on the variance of the two sets and their correlation, therefore maximizing the covariance also maximizes the other three.

$$\text{Cov}(t_a, u_a) = \text{Corr}(t_a, u_a) \times \sqrt{t_a^T t_a} \times \sqrt{u_a^T u_a} \tag{2.22}$$

One of the main advantages of using PLS over PCR is the fact that only one model is built for all the M variables of Y, instead of building M different models when using MLR or PCR. Moreover, PLS is more efficient than PCR due to the fact that it exploits the correlation between the columns of the score matrix T with respect to the columns in Y, thus increasing the captured variance and reducing the number of required principal components.
The relationship between the scores T and the data X is mediated by the $[K \times A]$ matrix

R which contains the regression coefficients. The R matrix is used instead of the W one because the latter represents the relationship between the scores and the deflated version of the X matrix. It can be shown that [2.23] holds between R and W is:

$$R = W(P^T W)^{-1} \tag{2.23}$$

The generally used algorithm to compute the loadings and the scores for PLS is the NIPALS one.
The steps are:

1. Selects a column from Y_a as the initial estimate for u_a

2. Iteratively regress each column of X_a onto u_a [2.24], the regression coefficients are stored as entries of w_a
$$w_a = \frac{1}{u_a^T u_a} \cdot X_a^T u_a \tag{2.24}$$
Normalize the vector

3. Iteratively regress each row of X_a onto w_a [2.25], the regression coefficients are stored as entries of t_a
$$t_a = \frac{1}{w_a^T w_a} \cdot X_a w_a \tag{2.25}$$
Normalize the vector

4. Iteratively regress each column of Y_a onto t_a [2.26], the regression coefficients are stored as entries of c_a
$$c_a = \frac{1}{t_a^T t_a} \cdot Y_a^T t_a \tag{2.26}$$
Normalize the vector

5. Iteratively regress each row of Y_a onto c_a [2.27], the regression coefficients are stored as entries of u_a
$$u_a = \frac{1}{c_a^T c_a} \cdot Y_a c_a \tag{2.27}$$
Normalize the vector

6. Deflate both the X and Y matrices to remove the already explained variability by the a

component

$$E_a = X_a - t_a p_a^T \quad \text{and} \quad F_a = Y_a - t_a c_a^T$$
$$X_{a+1} = E_a \quad \text{and} \quad Y_{a+1} = F_a$$

Where:

$$p_a = \frac{1}{t_a^T t_a} \cdot X_a^T t_a \quad \text{and} \quad u_a = \frac{1}{c_a^T c_a} \cdot Y_a c_a \tag{2.28}$$

7. Increment a and start again from step 2 until the wanted number of PCs is obtained

All the steps are resumed in figure [2.9], each arrow is a LS regression.

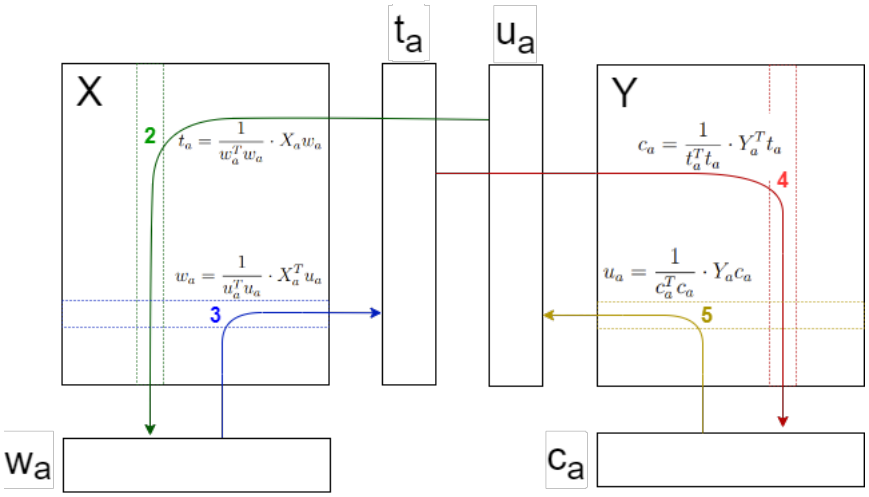


Figure 2.9: Graphical representation of the NIPALS algorithm, each arrow corresponds to a least square regression

2.1.6 Locally Weighted Regression (LWR)

Locally Weighted Regression is a non-parametric regression technique generally used to handle non linear problems. The method is based on the hypothesis that each y value is generated through a smooth function of the independent variables [2.29].

$$y_i = g(x_i) + \epsilon_i \tag{2.29}$$

where ϵ is a independent normal variable with mean 0 and variance σ^2.
The method estimates the function $\hat{g}(x)$ using only $q = 0, 1, ..., N$ samples from the calibration set whose distance from the reference point x_i is smaller than an hyperparameter set threshold. Generally the chosen metric is the Euclidean distance or the Mahalanobis one [2.30] applied onto the principal components space found by applying PCA.

$$D_M = \sqrt{(x-\mu)S^{-1}(x-\mu)} \tag{2.30}$$

where μ is the mean value and S is the covariance matrix.
The Mahalobis distance is preferred over the euclidean one since it also takes into account the correlation between the variables, as a matter of fact if the covariance matrix is equal to the identity matrix then the Mahalobis distance corresponds to the Euclidean one. The more similar the samples are the more weights they are given during the model computation, the used kernels are various and different, namely:

Kernel name	Function
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}$
Epanechnikov	$\frac{3}{4}(1-u^2)$
Tricube	$\frac{70}{81}(1-\lvert u\rvert^3)^3$

Table 2.1: Table of the generally used kernels used in LWR

Where u is the ratio between the distance of the near points x_i to x and their maximum value.
For each subset of similar samples a local model is built, generally PCR or PLS are used and the number of principal components used to regress can either be identified on the global dataset or on the local subset.
The local behaviour of the model helps adapting it to non linear conditions by approximate, with a piecewise linearization, the dynamic of the output. The concept is better explained graphically Fig.[2.10]

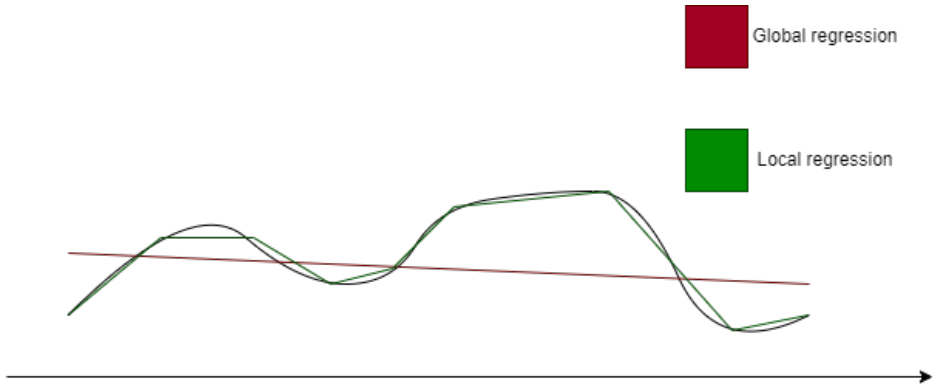


Figure 2.10: Graphical difference between local and global regression, in the global one although the sum of squares is minimized it is still high

2.1.7 Understanding scores and loadings plots

Both PCA and PLS are useful tools to analyze data and extract the main features. The output of both the analyses returns the scores and loadings which explain the same concept, but in a different way. The first aims at explaining at best a single dataset, while the second aims at explaining the correlation between the two.
Remember that the scores are the distance between the origin and the projection of the data onto the new axis and mathematically are computed as the linear combination between the data and the loadings, the latter effectively correspond to the weights assigned to each variable. In the scores plot are represented the observations in function of the scores on each principal components. The data points with a similar scores are close to each other meaning that the corresponding observations are similar, moreover if the points are close to the origin of the plot it means that the observations are close to the average. Knowing this information can be useful to cluster the data by building a confidence ellipse which is the region into which most of the distribution, generally 95%, of the data with the same properties can be found. In Fig.[2.11] is shown an example of a score plot based on a dataset of 36 observations, it is also shown how the data can be clustered in three different groups.

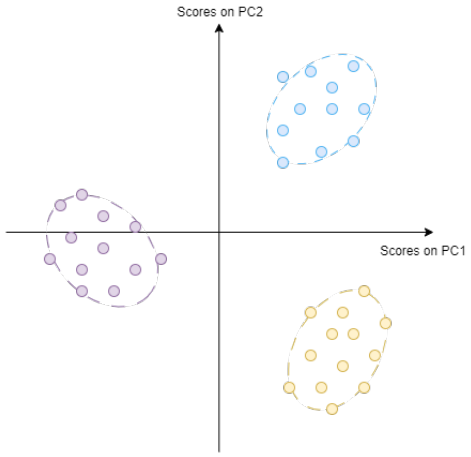


Figure 2.11: Example of a scores plot with 36 observation and 3 clusters

In the loadings plot are represented the variables in function of the loading contributions. The variables with similar loadings are close to each other meaning they have the same importance and a positive correlation exists, on the other hand when on opposite dials a negative correlation exists. It is often useful to compare both the scores plot and the loadings plot to understand the properties of a single observation. In Fig.[2.12] is shown an example of a loading plot based on a dataset of 5 variables.

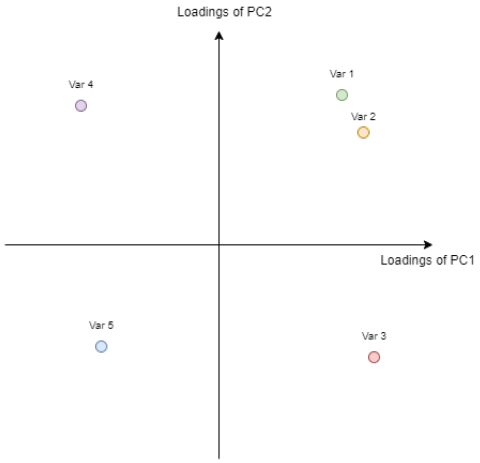


Figure 2.12: Example of a loading plot with 5 variables

2.2 Satisfaction Test

Environmental wellness is defined as occupying pleasant, stimulating environments that support well-being. It is an ensemble of various elements which influence it, namely the thermal comfort, clothing insulation, activity performed, air ventilation and relative humidity. The most important one is the thermal comfort.
Thermal comfort is the condition of mind that expresses satisfaction with the thermal environment, it is assessed by subjective evaluation and is a critical factor both for the health status and the mental well-being of an individual, particularly for those working in closed spaces. Thermal neutrality is maintained when the heat generated by human metabolism is allowed to dissipate, thus maintaining thermal equilibrium with the surroundings and is influenced by various factors, namely:

Factors	Measure Unit
Metabolic Rate	$1\ MET = 1\ kcal \cdot kg^{-1} \cdot h^{-1}$
Clothing Insulations	$1\ clo = 0.155\ K \cdot m^2 \cdot W^{-1}$
Air Temperature	$1\ °C = 1\ K - 273,15$
Mean Radiant Temperature	$1\ °C = 1\ K - 273,15$
Air Speed	$1\ kn = 1.852\ km \cdot h^{-1}$
Relative Humidity	adimensional

Table 2.2: Factors influencing thermal neutrality

Generally two metrics can be used to evaluate the environmental wellness, the first is the overall satisfaction in a scale from 1 (not satisfied) to 10 (perfectly satisfied) which is self-assessed through a questionnaire. The second one, mainly used to assess thermal comfort is the predicted mean vote (PMV), it is the mean response of a large group of people on a 7-point thermal sensation scale, from +3 (hot) to -3 (cold), where 0 is neutral. The metric is designed for fully mechanically ventilated buildings and is determined in accordance with ANSI/ASHRAE [1] Standard 55 and UNI EN ISO 7730.
A model for thermal comfort was formulated in 1970 by Fanger [29] and standardized in both the above mentioned regulations, the PMV is determined from the heat balance equation [2.31] of the human being in steady-state with the environment:

$$PMV = (q_M - q_j)[0.303 * exp(-0.036M) + 0.028] \tag{2.31}$$

where M is the metabolic rate, q_M is the heat production and depends on the person's activity

[1] American National Standards Institute and American Society of Heating, Refrigerating and Air-Conditioning Engineers

and q_j is the human heat loss which takes into account the clothing insulation. The metabolic rate [2.32] can be defined as the product between the standard metabolic unit (MET) and the resting metabolic rate RMR.

$$M = MET * RMR \tag{2.32}$$

Based on the Harris-Benedict equation, the RMR differs between men and women:

$$\text{Men RMW } [W] = \frac{1.163}{24}(66.47 + 13.75W_b + 500H_b - 6.75 \cdot age) \tag{2.33}$$

$$\text{Women RMW } [W] = \frac{1.163}{24}(665.09 + 9.56W_b + 184H_b - 4.67 \cdot age) \tag{2.34}$$

For both ASHRAE Standard 55 and UNI EN ISO 7730 PMV should be in range between ± 0.5, corresponding to a predicted percentage of dissatisfied (PPD) below 10%, the index can be evaluated on the basis of PMV as:

$$PPD = 100 - 95 \cdot exp(-0.03353 \cdot PMV^4 - 0.2179 \cdot PMV^2) \tag{2.35}$$

PPD corresponds to the percentage of people in a room who do not feel comfortable with their surroundings. The relation between PPD and PMV is showed in Fig.[2.13]

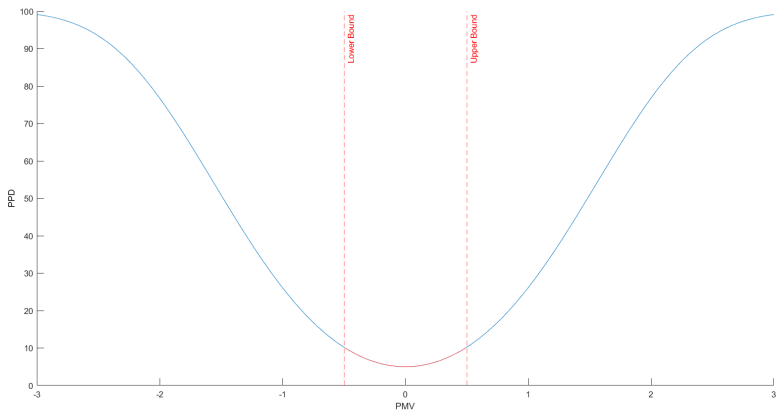


Figure 2.13: Plot of PPD with respect to PMV, the rule of thumbs states that for a room to be comfortable to thevels of PPD should be below 10% corresponding to PMV between ± 0.5

Chapter 3

EXPERIMENTAL SETUP AND ANALYSIS

3.1 Protocol

The sensors are placed in three rooms of the University of Bologna - Cesena Campus, one is placed in a smaller room with a small amount of people while the other two are placed in a bigger space with a higher density of people. The first serves as control group, the second and third as the main target.
The initial X matrix is comprehensive of 3024 observations and 11 variables. The variables are shown below with the respective measure unit[1].

Index	Variables	Measure Unit
1	Temperature	[°C]
2	Pressure	[bar]
3	Humidity	[%]
4	CO_2 Concentration	[ppm]
5	Indoor Air Quality	[]
6	Static Indoor Air Quality	[]
7	VOC Concentration	[ppm]
8	People Density	[# of People / m^2]
9	External Temperature	[°C]
10	External Humidity	[%]
11	Time	[min]

[1]Empty brackets indicate an adimensional measure unit

The data regarding the external temperature and external humidity is gathered trough the free VisualCrossing database [30], the others via the Taunito sensors.
The data gathered from the Taunito sensors are continuously sampled every 20 minutes and uploaded on the online app every 2 hours. From the web interface the data is downloaded in .csv format Fig.[3.1].

Figure 3.1: Interface of the Taunito sensors

The external data is downloaded on the Visual Crossing database in .csv format, only data related to the temperature and humidity are retained. The data from the VS database are sampled each hour.
The number of people in the rooms are manually evaluated every day from 09:00 up to 18:00, from Monday up to Friday. The measures are done at four different time instants: 09:00, 12:00, 15:00 and 18:00.
The Y matrix contains either the satisfaction scores or the PMV, both are acquired through the response to a questionnaire. The test is performed on a group of around 20 people for the control room, 50 people for the first objective room and 30 for the second one. The test is built on MS Forms and automatically shared through a flux created on MS PowerAutomate on a MS Teams Channel with all the voluntary participants in it. The scores are measured at the same above mentioned time instants. The flow is illustrated in Fig.[3.2]

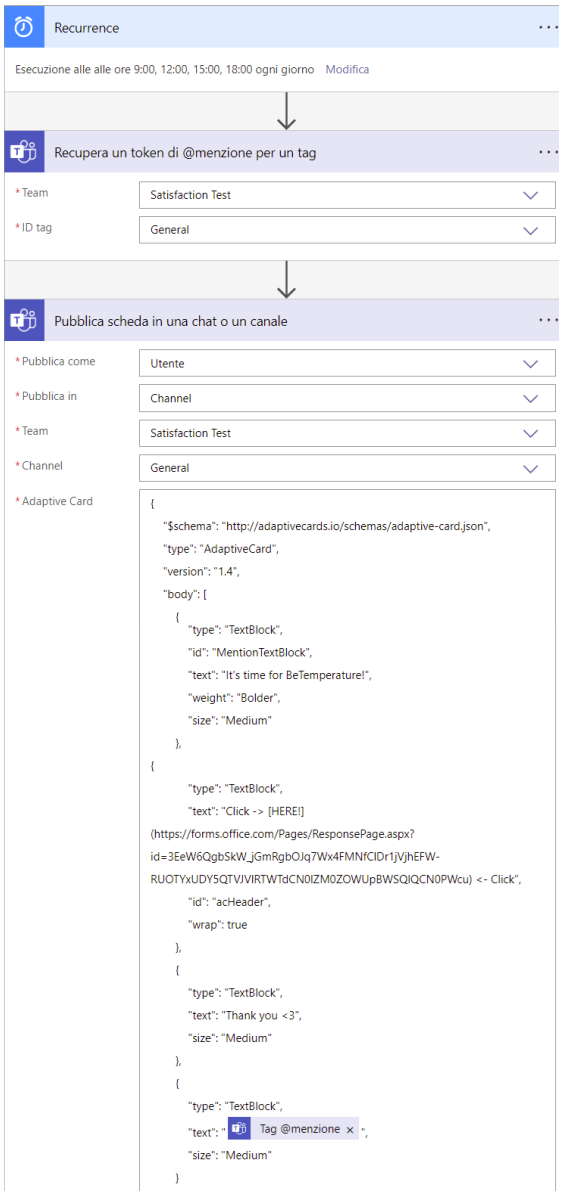


Figure 3.2: Flux created through PowerAutomate which sends at specific time instants a reminder for the form to all the voluntary participants

The table resumes the ways to acquire data:

Variable	Acquisition Method	Sampling Time
Temperature	Tauanito Sensor	every 20 min
Humidity	Tauanito Sensor	every 20 min
Pressure	Tauanito Sensor	every 20 min
CO_2 Concentration	Tauanito Sensor	every 20 min
Index of Air Quality	Tauanito Sensor	every 20 min
VOC Concentration	Tauanito Sensor	every 20 min
External Temperature	Visual Crossing Database	every 1 h
External Humidity	Visual Crossing Database	every 1 h
People Density	Manual counting	09:00/12:00/15:00/18:00
Time	Tauanito Sensor	every 20 min
Predicted Mean Value	MS Form	09:00/12:00/15:00/18:00
Satisfaction	MS Form	09:00/12:00/15:00/18:00

The data is gathered for a period of 21 days, starting from the 23^{th} Nov 2023 up to 13^{th} Dec 2023, the samples relative to Sundays, Saturdays and holidays (8^{th} Dec.) have been discarded due to the lack of people in the rooms at the time of acquisition.

3.2 Analysis Setup

The analysis of data is performed in the MATLAB environment, using the PLS toolbox, property of Eigenvector Research, Inc. which includes many machine learning approaches to build a predictive and classification model.

The data is pre-processed to remove possible artifacts caused by the involuntary touches of the device, outliers are removed using the 3σ rule, therefore all the observations where at least one of the variables is out of the 99% range is removed. To address the different sampling times each observation is linearly re-sampled under the hypothesis of a local slow dynamic allowing comparison between the observations.

The elaborated X dataset, free of outliers and time selected, is comprehensive of 769 observations, similarly the Y dataset.

Two classes groups are defined, the first associates each observation to the location while the second is relative to the different times of the day as follows:

Class name	Time interval
Morning	09:00 up to 12:00
Lunch Time	12:00 up to 14:00
Afternoon	14:00 up to 18:00

In Fig.[3.3] and Fig.[3.4] are shown the plots of the variables over time.

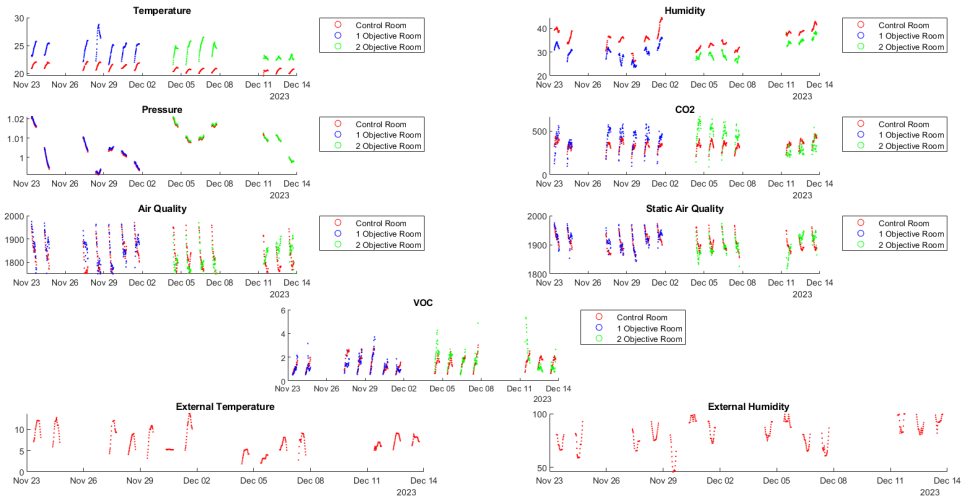


Figure 3.3: Time plots of some of the variables distinct based on device class

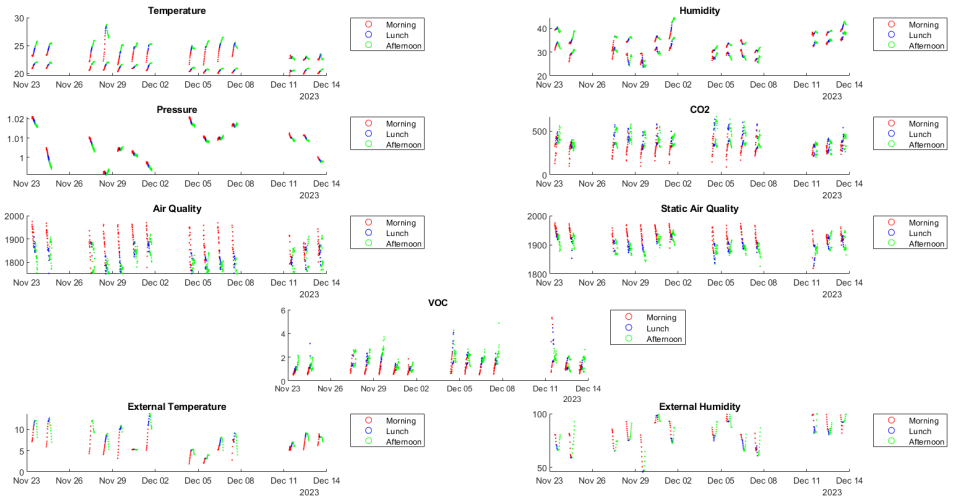


Figure 3.4: Time plots of some of the variables distinct based on time of the day

The training and test set are randomly built with a ratio of respectively 0.8 and 0.2 with respect to the size of the entire dataset.

For all the different analysis run onto the data cross-validation is performed using a Venetian blind algorithm with 10 splits and unitary thickness. More generally for a size N dataset applying G splits with thickness T each entry at index $(k-1)\cdot G+t+(m-1)\cdot T$ is assigned

the same cross-validation group, with k in $range(1, N\backslash G)$ and m in $range(1, G\backslash T)$. This is better explained graphically Fig.[3.5]

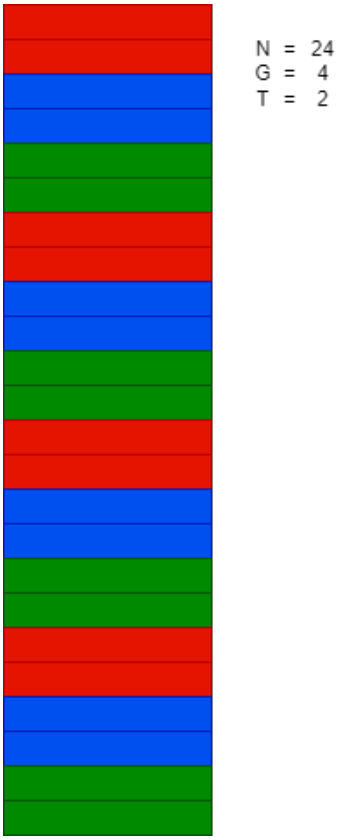


Figure 3.5: Graphical example of Venetian blind cross-validation using 24 samples, 4 groups and thickness equal to 2

Following are described the analysis procedures for the choice of the pre-processing to use onto the calibration data, in particular are showed the metrics obtained when no pre-processing is involved and when either mean centering and autoscaling are used:

- **PCA SETUP:**
  The use of PCA is to look for both a correlation between the physical quantities and the satisfaction level and both for a statistical difference between the locations. The first goal is achieved via the loading plots, the second via the score plot.
  Prior to the calibration step, the data is preprocessed to bring forth the features. In the table below are showed the metrics used to quantify the efficiency of the method such as RMSEC (calibration) and RMSECV (cross-validation) with respect to the different preprocessing techniques used.

Preprocessing	# PC	Captured Variance	RMSEC	RMSECV
None	4	100%	3.89	228.4
Mean Center	4	99.70%	3.11	138.90
Autoscale	4	79.53%	0.45	0.67

Table 3.1: Table with the metrics for each pre-processing obtained through PCA

As showed by the metrics the best preprocessing technique is autoscale.

- **SIMCA setup:**
  The use of SIMCA is to assign each observation to a class group. The model is built on the class sets identifying the location and for each a PCA model is built. Below are showed the metrics [3.2] and confusion matrices [3.3] obtained when varying the pre-processing.

Preprocess	$\#PC$	RMSE	RMSECV	P	F_1	Misclassifaction Error
None	3	5.12	51.97	0.73	0.84	0.18
	6	0.41	86.02	0.77	0.72	0.14
	8	0.14	55.01	1	0.50	0.15
Mean	1	17.18	43.32	0.71	0.81	0.21
	3	4.37	186.8	0.56	0.57	0.23
	1	36.91	14.05	0.73	0.31	0.20
Autoscale	4	0.43	0.69	1	0.95	0.02
	6	0.24	0.74	0.92	0.91	0.05
	6	0.22	0.50	0.89	0.97	0.02

Table 3.2: Table with the metrics for each pre-processing, for each row corresponds the correspective class

No preprocess involved				
	TPR	FPR	TNR	FNR
Class 1	1	0.38	0.62	0
Class 2	0.67	0.08	0.92	0.33
Class 3	0.33	0	1	0.67

(a) None

Preprocess = Mean Center				
	TPR	FPR	TNR	FNR
Class 1	0.94	0.38	0.62	0.06
Class 2	0.58	0.17	0.83	0.42
Class 3	0.19	0.03	0.97	0.81

(b) Mean Center

Preprocess = Autoscale				
	TPR	FPR	TNR	FNR
Class 1	0.94	0	1	0.06
Class 2	0.92	0.05	0.95	0.08
Class 3	0.98	0.03	0.97	0.02

(c) Autoscale

Table 3.3: Confusion matrices varying the preprocess

As showed by the metrics and confusion matrices the best preprocessing technique is autoscale.

- **PLS setup:**
  The use of PLS is to both look for a correlation between the physical quantities and the satisfaction level and both for predicting the satisfaction level through the variables. Below are shown the different parameters with respect to the different preprocessing technique

Preprocessing	# PC	X Variance Explained	Y Variance Explained	RMSEC	RMSECV
None	4	100%	30.72%	0.95	0.96
Mean Center	4	99.61%	34.95%	0.92	0.93
Autoscale	4	67.86%	70.69%	0.62	0.66

Table 3.4: Table with the metrics for each pre-processing obtained through PLS

As showed by the metrics the best preprocessing technique is autoscale.

- **LWR setup:**
  The use of LWR is to optimize the regression since it best handles non linear problems. The model is built on 10 local points, below are showed the parameters used when varying the pre-processing.

Preprocessing	# PC	Captured Variance	RMSEC	RMSECV
None	4	100%	0.26	1.64
Mean Center	4	98.42%	0.18	1.76
Autoscale	4	80.01%	0.07	0.39

Table 3.5: Table with the metrics for each pre-processing obtained through LWR

- **PLS setup for prediction of optimal temperature:**
  To predict the optimal temperature for which the thermal comfort is maximized a new dataset is built comprehensive of all the environmental quantities with the exception of temperature replaced with the PMV scores. The choice of using the PMV scores instead of the satisfaction ones is due to the inherent non linear relationship between satisfaction and temperature, as an example low scores can either represent a feeling of excessive coldness or hotness.
  A regression model is built to predict the temperature, three preprocessing algorithm are used and the efficiency metrics are shown below.

Preprocessing	# PC	X Captured Variance	Y Captured Variance	RMSEC	RMSECV
None	5	100%	54.02%	1.33	1.33
Mean Center	5	99.82%	60.76%	1.23	1.23
Autoscale	6	85.84%	89.63%	0.63	0.64

Table 3.6: Table with the metrics for each pre-processing obtained through PLS

As showed by the metrics the best preprocessing technique is autoscale.
Once the model is built the regression vector (R) is extracted which mediates the linear relationship between the input and output of the model, for the temperature to be optimal the PMV scores are all equal to 0, mathematically [3.7] it can be translated as:

$$\begin{aligned} \text{Temp} &= R_1 \cdot Var_1 + R_2 \cdot Var_2 + ... + R_N \cdot PMV \\ \text{Opt. Temp} &= R_1 \cdot Var_1 + R_2 \cdot Var_2 + ... + R_N \cdot 0 \end{aligned}$$

Table 3.7: Mathematical way to obtain the optimal temperature

- **LWR setup for prediction of optimal temperature:**
  The same goal of the previous step is done, with a different approach, using LWR. Two datasets are built, the first called X_{ctr} only contains the data for which the satisfaction scores are above a set threshold (= 7.5), while X_{obj} contains the remaining. The X_{ctr}

dataset is used to calibrate a new LWR model using 10 local points, following are shown the metrics measured varying the pre-process:

Pre-processing	# PC	Captured Variance	RMSEC	RMSECV
None	3	100%	0.12	0.37
Mean Center	4	99.81%	0.05	0.38
Autoscale	6	84.46%	0.004	0.34

Table 3.8: Metrics obtained through LWR by varying the pre-processing

As showed by the metrics the best pre-processing technique is autoscale.
The observations belonging to X_{obj} are projected onto the score space via the previously built model. All the scores extracted through X_{ctr} are used to compute the reference point as their mean. The algorithm moves all the objective points towards the reference one in order to minimize the euclidean distance between the two.

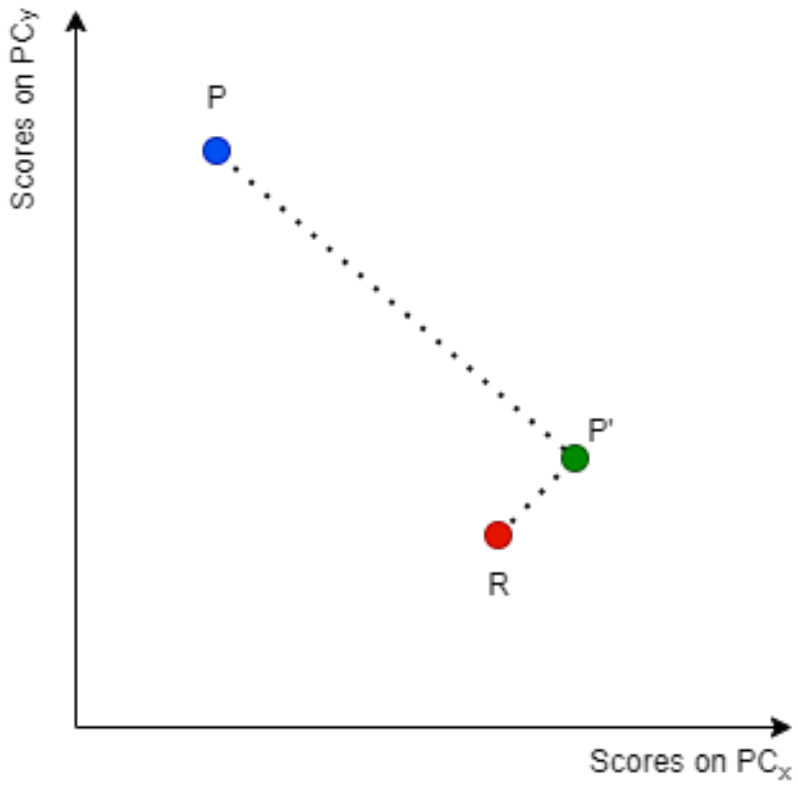


Figure 3.6: Graphical interpretation of the algorithm, the objective point P moves in a straight line to reach point P' which minimizes the distance from the reference point

Mathematically it is translated as in [3.9] whereas P is the generic point in the score

plot, P' is the modified point in the score plot, V_j is the $j-th$ variable of the observation, L_j is the $j-th$ loading and Δ is the vector of the variable changes.

$$
\begin{aligned}
P &= V_1 \cdot L_1 + V_2 \cdot L_2 + ... + V_K \cdot L_K \\
&= V \cdot L \\
P' &= (V_1 + \Delta V_1) \cdot L_1 + (V_2 + \Delta V_2) \cdot L_2 + ... + (V_K + \Delta V_K) \cdot L_K \\
&= V \cdot L + \Delta \cdot L \\
&= P + \Delta \cdot L \\
dist(R, P')^2 &= \sum_{a=1}^{A} (R_j - P'_j)^2 \\
&= \sum_{a=1}^{A} (R_j - P_j - \Delta V_j \cdot L_j)^2 \\
&= \sum_{a=1}^{A} (R_j - P_j)^2 + \sum_{a=1}^{A} (\Delta V_j \cdot L_j)^2 - 2 \sum_{a=1}^{A} (R_j - P_j - \Delta V_j \cdot L_j) \\
&= dist(R, P)^2 + ||\Delta \cdot L||_2^2 - 2 \sum_{a=1}^{A} (R_a - P_a - \Delta \cdot L_a) \\
\hat{\Delta} &\doteq \underset{\Delta}{argmin}\ dist(R, P')
\end{aligned}
$$

Table 3.9: First approach by minimizing the distance between the two points

A similar approach is to find the Δ values for which the vector PP' and RP' are orthogonal. Mathematically this translates as in [3.10]:

$$
\begin{aligned}
PP' &= P' - P \\
&= \Delta \cdot L \\
RP' &= P' - R \\
&= RP + \Delta \cdot L \\
\hat{\Delta} &\doteq (\hat{\Delta} \cdot L) \cdot (RP + \hat{\Delta} \cdot L)^T = 0
\end{aligned}
$$

Table 3.10: Second approach by finding the Δ values for which the two vectors are orthogonal

The optimal temperature is then computed by adding the first entry of the Δ vector to the measured one.

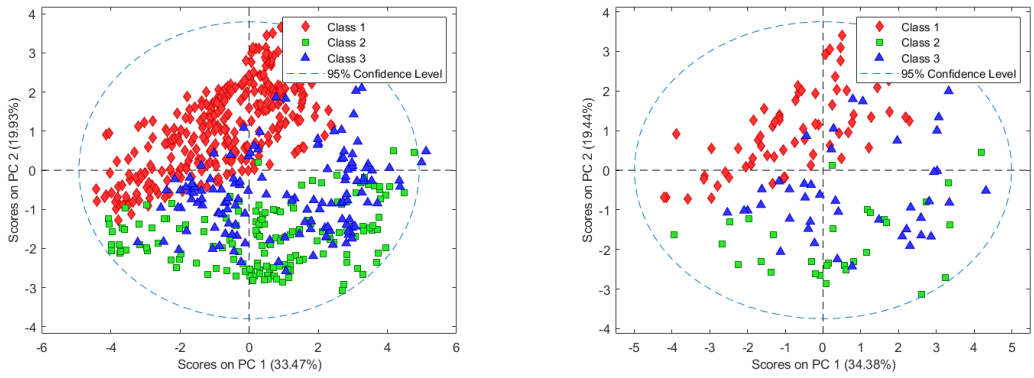
Chapter 4

RESULTS AND DISCUSSION

In this chapter are shown the results of the analysis which are later discussed to extract the meaningful info.

4.1 Results

The results of the analysis are obtained from the test sets applied onto the calibrated models discussed in the previous chapter. Although methods to automatically remove outliers exists, it was chosen to keep them both for completeness and for checking robustness of the model. In the following it is chosen to denote the control room as ”Class 1”, while the other two objective ones as ”Class 2” and ”Class 3” respectively.
The decomposition analysis (PCA) was executed using 4 principal components and applying autoscale onto the data. Both the calibration and test scores plot Fig.[4.1] and loading plot Fig.[4.2] are shown below.



(a) Score plot obtained on the calibration set

(b) Score plot obtained on the test set

Figure 4.1: Scores plot of the test set applied onto the PCA model

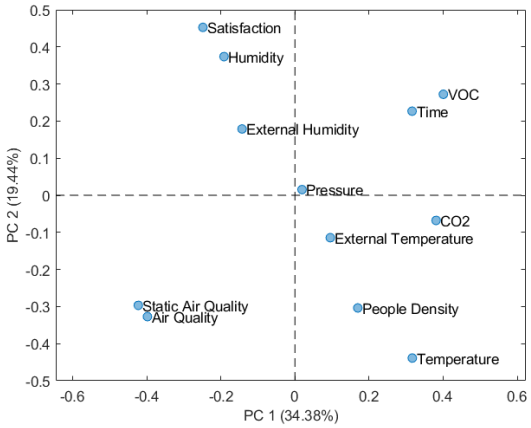


Figure 4.2: Loading plot of the test set applied onto the PCA model

The classification algorithm (SIMCA) was executed using 4,6 and 6 principal components respectively for the first, second and third class and using autoscale onto the data.

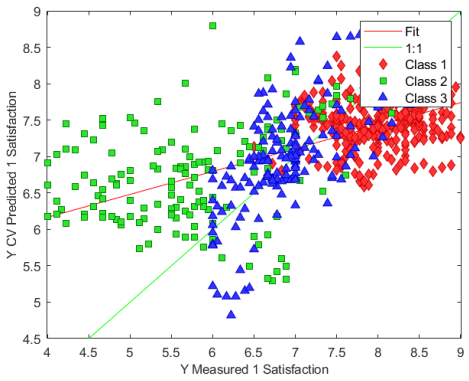
	P	F_1	Misclassification Error
Class 1	0.97	0.98	0.01
Class 2	0.80	0.88	0.06
Class 3	1.00	0.88	0.06

Table 4.1: Metrics of the test set applied onto the SIMCA model

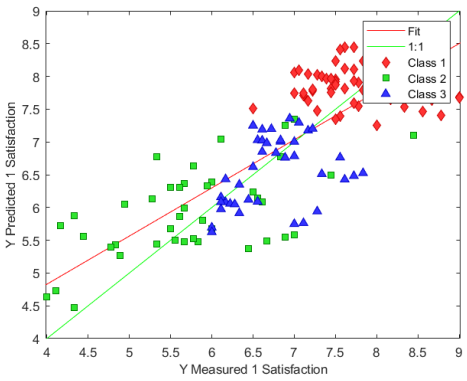
	TPR	FPR	TNR	FNR
Class 1	1.00	0.02	0.97	0.00
Class 2	0.97	0.07	0.93	0.03
Class 3	0.79	0.00	1.00	0.20

Table 4.2: Confusion matrix of the test set applied onto the SIMCA model

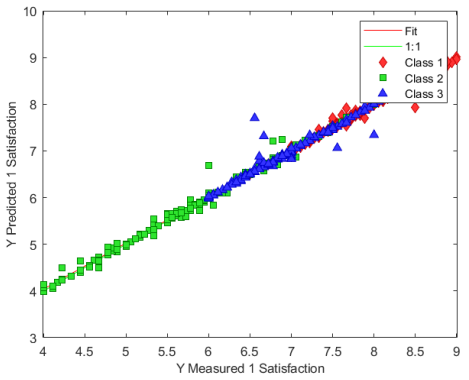
The regression analysis (PLS and LWR) onto the satisfaction scores was executed using 4 principal components (on 10 local points for LWR) and applying autoscale onto the data.



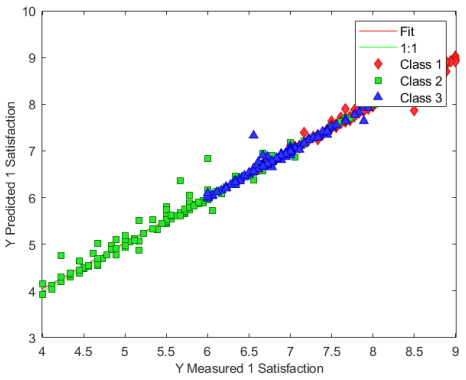
(a) Regression curve of the calibration set applied onto the PLS model



(b) Regression curve of the test set applied onto the PLS model



(c) Regression curve of the calibration set applied onto the LWR model



(d) Regression curve of the test set applied onto the LWR model

The regression analysis (PLS) onto the optimal temperature was executed using 6 principal components and applying autoscale onto the data Fig.[4.4].

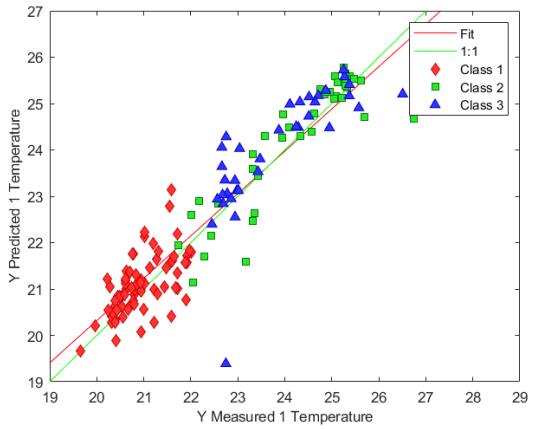


Figure 4.4: Regression curve of the test set applied onto the PLS model for the prediction of optimal temperature

The predicted temperature is graphically compared with the actual measured one as in Fig.[4.5]

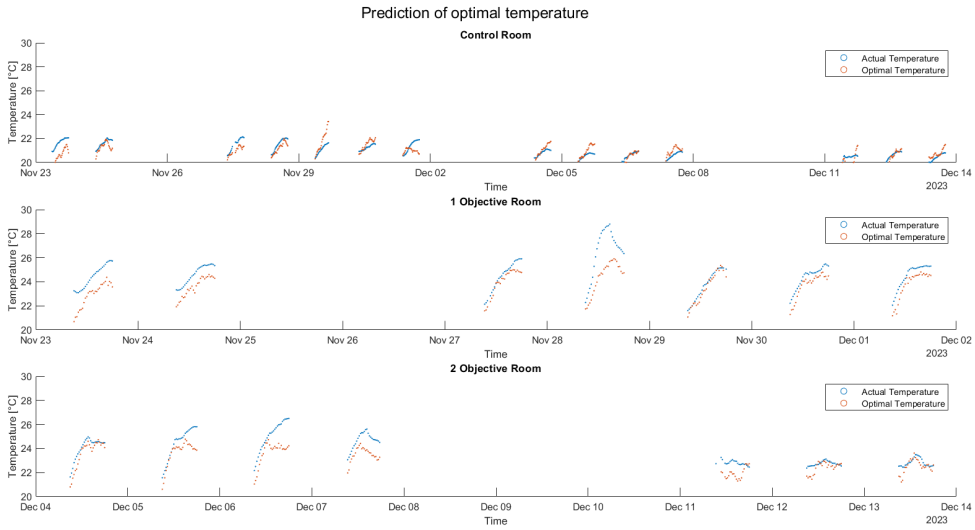


Figure 4.5: Time plot prediction of the optimal temperature using the PLS method in the three rooms

To counter-check the results a new dataset with the same features, with the exception of temperature replaced with the predicted one, is used to regress onto the PMV. The predicted PMV are used to evaluate the PPD according to the Fanger's model. The results are shown in Fig.[4.6]

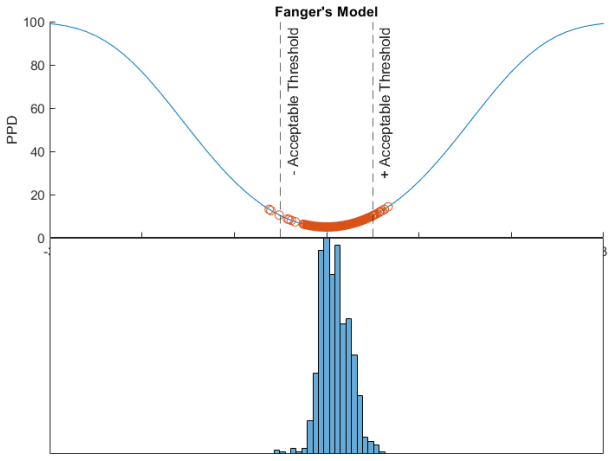
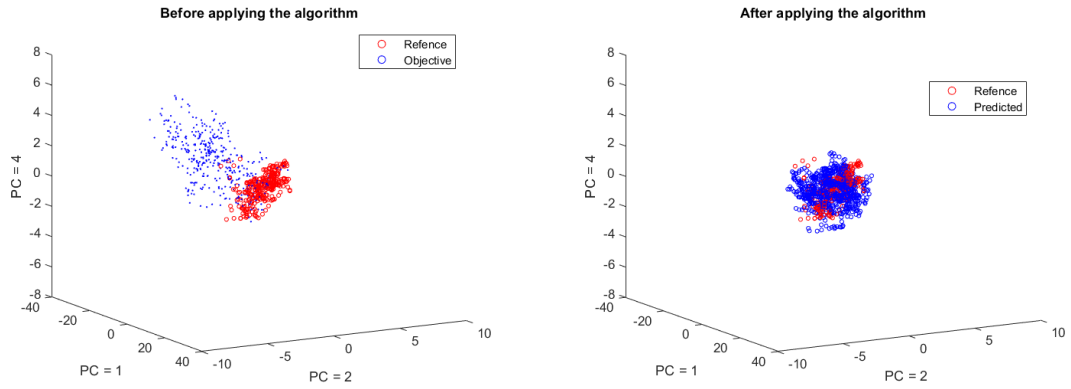


Figure 4.6: Fanger's model applied onto the predicted PMV, the results show a percentage of satisfaction of 98.82%

If the PMV values fall in the range of [-0.5,+0.5] (or equally for PPD $\leq 10\%$) then the subjects are satisfied with the environment. The results show a satisfaction percentage of 98.82%

.

The regression analysis (LWR) onto the optimal temperature was executed using 6 principal components onto 10 local points and applying autoscale onto the data. The scores plot showing the observations before and after applying the algorithm are shown in Fig.[4.7]



(a) Score plot before applying the algorithm (b) Score plot after applying the algorithm

Figure 4.7: Scores plot before and after the algorithm. In red are shown the points for which the satisfaction is above the threshold and used for the computation of the reference point, in blue the objective ones to improve

The optimal predicted temperature for which the objective points are closer to the reference one is shown in Fig.[4.8]

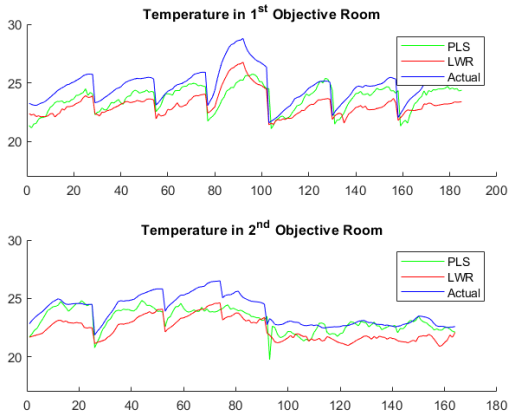


Figure 4.8: Graph of the optimal predicted temperature both using the previous algorithm with PLS (green) and the one using LWR (red) with respect to the actual temperature

To assess the effectiveness of the algorithm a new dataset which replaces the temperature of the original dataset with the new predicted one is used to predict the satisfaction scores.

The success rate is evaluated as in [4.1] resulting in a success rate of 75.68%.

$$Succ[\%] = \frac{len(Y_{new} \geq Y_{old})}{len(Y_{old})} \cdot 100 \tag{4.1}$$

4.2 Discussion

After collecting the results several key findings emerged.
Firstly, from the correlation analysis and with reference to the loading plot Fig.[4.2] there exists, as expected, a correlation between air quality and VOC since the first is the result of an algorithm mainly computed on the second. Most importantly there exists a clear negative correlation between people' satisfaction and temperature, the nature of this relation, either linear or non-linear, is further investigated in the regression analysis.
From the classification task and with reference to the confusion matrix Fig.[4.2], it is clear how machine learning approaches are able to accurately distinguish between acceptable and non-acceptable rooms. The classification can also be used to assign a reference group as described in EN ISO 7730 to speed up the ranking of the rooms since in the current state in order to classify a room a special software is required which uses many variables, some of which hard to measure.
The first regression task is aimed at predicting the satisfaction scores, but it also serves as a way to identify the relation between satisfaction and the variables, mainly temperature. As shown in the results Fig.[4.3a] and Fig.[4.3b] the PLS method fails to accurately predict the scores being a linear approach, on the other hand, LWR performs better in the task Fig.[4.3c] and Fig.[4.3d]. This knowledge confirms the hypothesis of non linearity between the input and output of the model due to the inherent nature of the satisfaction scores, for this reason in the following task the PMV scores are used to predict the temperature.
The second regression task is aimed at predicting the optimal temperature for which the percentage of dissatisfied people is minimized, or equally, for which the average satisfaction is maximized. In the first algorithm, comparing the measured temperature and the optimal predicted one Fig.[4.5] is clear how in the control room there is no statistical difference ($P > 0.05$), while in the other two objective rooms is predicted a lower temperature than the actual one ($P < 0.05$), this concept is validated by the answers of the occasional interviews to the participants in the project. The created model can be applied onto a micro-controller connected to a thermostat in order to keep the optimal temperature.
The second algorithm uses a non linear approach to reach the same goal, also in this case the predicted temperature is less than the actual one in the objective rooms Fig.[4.8]. The advantage of using this method is the possibility to control all the other variables instead of just one, thus providing more flexibility and control over the environment.
The work showed how machine learning approaches can help to have a better insight on the understanding of data and how this can be exploited to improve people lives under the psychophysical aspect. Although the study shows the possibility to do so it is important to remember

the limitations of machine learning, the biggest is the necessity of training a model specifically for the objective environment thus requiring time before the effective implementation. Future studies are required to address this problem and to effectively validate the results in a more general context, especially for the prediction of the optimal temperature.

Bibliography

[1] Yoichi Chida and Andrew Steptoe. “Positive psychological well-being and mortality: a quantitative review of prospective observational studies”. In: *Psychosomatic medicine* 70.7 (2008), pp. 741–756.

[2] WHOQoL Group. “Study protocol for the World Health Organization project to develop a Quality of Life assessment instrument (WHOQOL)”. In: *Quality of life Research* 2 (1993), pp. 153–159.

[3] Wikipedia. *Wellness (alternative medicine) — Wikipedia, The Free Encyclopedia*. `http://en.wikipedia.org/w/index.php?title=Wellness\%20(alternative\%20medicine)&oldid=1199686666`. 2024.

[4] Bill Hettler. “Wellness promotion on a university campus”. In: *Family & community health* 3.1 (1980), pp. 77–95.

[5] Debbie L Stoewen. “Dimensions of wellness: Change your habits, change your life”. In: *The Canadian veterinary journal* 58.8 (2017), p. 861.

[6] CHARLENE W BAYER. “Indoor Environmental Quality and Health Improvement, Evidence-Based Design for”. In: (2013).

[7] National Institutes of Health. *Common Fund*.

[8] CNX OpenStax.

[9] Benjamin Hing et al. “Chronic social stress induces DNA methylation changes at an evolutionary conserved intergenic region in chromosome X”. In: *Epigenetics* 13.6 (2018), pp. 627–641.

[10] Bridget R Mueller and Tracy L Bale. “Sex-specific programming of offspring emotionality after stress early in pregnancy”. In: *Journal of Neuroscience* 28.36 (2008), pp. 9055–9065.

[11] Irina Bancos et al. “Evaluation of FKBP5 as a cortisol activity biomarker in patients with ACTH-dependent Cushing syndrome”. In: *Journal of Clinical & Translational Endocrinology* 24 (2021), p. 100256.

[12] Xiaojing Zhang et al. “Effects of exposure to carbon dioxide and bioeffluents on perceived air quality, self-assessed acute health symptoms, and cognitive performance”. In: *Indoor air* 27.1 (2017), pp. 47–64.

[13] Bowen Du et al. “Indoor CO2 concentrations and cognitive function: A critical review”. In: *Indoor air* 30.6 (2020), pp. 1067–1082.

[14] Gregory C Pratt et al. “An assessment of air toxics in Minnesota.” In: *Environmental Health Perspectives* 108.9 (2000), pp. 815–825.

[15] Mahbub Rashid and Craig Zimring. “A review of the empirical literature on the relationships between indoor environment and stress in health care and office settings: Problems and prospects of sharing evidence”. In: *Environment and behavior* 40.2 (2008), pp. 151–190.

[16] Wikipedia. *Internet of things — Wikipedia, The Free Encyclopedia*. `http://en.wikipedia.org/w/index.php?title=Internet\%20of\%20things&oldid=1211504227`. 2024.

[17] Shivani Sharma, Sumit Kumar, et al. “A review on IoT: Protocols, architecture, technologies, application and research challenges”. In: *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE. 2020, pp. 559–564.

[18] Mary K. Pratt. *Top 12 most commonly used IoT protocols and standards*. July 2023. URL: `https://www.techtarget.com/iotagenda/tip/Top-12-most-commonly-used-IoT-protocols-and-standards`.

[19] Akira Umeda et al. “Recent insights into the measurement of carbon dioxide concentrations for clinical practice in respiratory medicine”. In: *Sensors* 21.16 (2021), p. 5636.

[20] Joaquim Bauxell, Mercè Vall-Llossera, and Hellen Gurgel. “Machine Learning Techniques Using Enviromental Data from Remote Sensing Applied to Modeling Dengue Risk in Brazil”. In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE. 2021, pp. 4608–4611.

[21] Emily Oldham and Hyojin Kim. “IEQ field investigation in high-performance, urban elementary schools”. In: *Atmosphere* 11.1 (2020), p. 81.

[22] Nishant Raj Kapoor et al. “A review on indoor environment quality of Indian school classrooms”. In: *Sustainability* 13.21 (2021), p. 11855.

[23] Sepideh Sadat Korsavi, Azadeh Montazami, and Dejan Mumovic. “The impact of indoor environment quality (IEQ) on school children’s overall comfort in the UK; a regression approach”. In: *Building and Environment* 185 (2020), p. 107309.

[24] Jiantao Weng et al. “Field Measurements and Analysis of Indoor Environment, Occupant Satisfaction, and Sick Building Syndrome in University Buildings in Hot Summer and Cold Winter Regions in China”. In: *International Journal of Environmental Research and Public Health* 20.1 (2022), p. 554.

[25] Steven Van Der Valk et al. “Sensor networks in workplaces: Correlating comfort and productivity”. In: *2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*. IEEE. 2015, pp. 1–6.

[26] Yuanchao Ma et al. “Daily mood assessment based on mobile phone sensing”. In: *2012 ninth international conference on wearable and implantable body sensor networks*. IEEE. 2012, pp. 142–147.

[27] Yadolah Dodge. *The Oxford dictionary of statistical terms*. OUP Oxford, 2003.

[28] Svante Wold. “Cross-validatory estimation of the number of components in factor and principal components models”. In: *Technometrics* 20.4 (1978), pp. 397–405.

[29] Poul O Fanger et al. “Thermal comfort. Analysis and applications in environmental engineering.” In: *Thermal comfort. Analysis and applications in environmental engineering.* (1970).

[30] URL: https://www.visualcrossing.com/.