buildings

Article

Analysis of Variables Affecting Indoor Thermal Comfort in Mediterranean Climates Using Machine Learning

Pablo Aparicio-Ruiz *, Elena Barbadilla-Martín, José Guadix and Julio Nevado

Organisation Engineering Group, School of Engineering, University of Seville, Camino de los Descubrimientos s/n, 41092 Seville, Spain; ebarbadilla@us.es (E.B.-M.); guadix@us.es (J.G.)
* Correspondence: pabloaparicio@us.es

Abstract: To improve the energy efficiency and performance of buildings, it is essential to understand the factors that influence indoor thermal comfort. Through an extensive analysis of various variables, actions can be developed to enhance the thermal sensation of the occupants, promoting sustainability and economic benefits in conditioning systems. This study identifies eight key variables: indoor air temperature, mean radiant temperature, indoor globe temperature, CO_2, age, outdoor temperature, indoor humidity, and the running mean temperature, which are relevant for predicting thermal comfort in Mediterranean office buildings. The proposed methodology effectively analyses the relevance of these variables, using five techniques and two different databases, Mediterranean climate buildings published by ASHRAE and a study conducted in Seville, Spain. The results indicate that the extended database to 21 variables improves the quality of the metrics by 5%, underscoring the importance of a comprehensive approach in the analysis. Among the evaluated techniques, random forest emerges as the most successful, offering superior performance in terms of accuracy and other metrics, and this method is highlighted as a technique that can be used to assist in the design and operation or control of a building's conditioning system or in tools that recommend adaptive measures to improve thermal comfort.

Keywords: building environment; thermal comfort; HVAC; machine learning

Citation: Aparicio-Ruiz, P.; Barbadilla-Martín, E.; Guadix, J.; Nevado, J. Analysis of Variables Affecting Indoor Thermal Comfort in Mediterranean Climates Using Machine Learning. *Buildings* **2023**, *13*, 2215. https://doi.org/10.3390/buildings13092215

Academic Editors: Diego Pablo Ruiz Padillo, María Dolores Martínez Aires and María Luisa de la Hoz Torres

Received: 20 July 2023
Revised: 21 August 2023
Accepted: 23 August 2023
Published: 30 August 2023

1. Introduction

Indoor thermal comfort research has considered a wide range of variables derived from both the interior and the exterior of buildings. These variables are used to predict the energy performance as well as the thermal comfort of buildings or the comfort experienced by their occupants. In order to achieve this goal, field studies usually use a Likert scale, consisting of five or seven points, to assess the thermal sensation, with it being expressed as: +3 hot, +2 warm, +1 slightly warm, 0 neutral, −1 slightly cool, −2 cool, and −3 cold. The response or vote refers to the thermal sensation produced by the transfer of heat between the body and the environment, although this may be linked to different variables and allows the percentage of dissatisfied people to be determined in a particular set of conditions.

Traditional approaches to thermal comfort determine a set of variables related to it. Fanger's static model is based on six variables that affect thermal comfort: air temperature, relative humidity, mean radiant temperature, air velocity, metabolic rate, and clothing insulation [1]. On the other hand, the adaptive model [2] maintains that the occupants of a building actively generate a thermal environment through corrective measures that allow them to restore their comfort; therefore, the model analyses the comfort temperature based on the indoor temperature and the running mean temperature. The latter includes in the model the effect of the outdoor temperature over the last few days to reflect the adaptation of users to the climate and its effect on the acceptance of an indoor temperature, called a comfort temperature. However, other comfort models applied to indoor and outdoor comfort can be found in the literature [3].

The thermal preferences of the users of a building are better investigated using a calised adaptive model consisting of specific country and regional data, since adaptive models differ according to the climate zone, country, and weather conditions. Therefore, the adaptive model ASHRAE-55:2020 considers data on global thermal comfort that represent all the significant climate areas, including the Tropics [4].

In addition, recent years have seen an increase in the use of machine learning (ML) and data mining techniques as tools to identify relationships that have gone unnoticed in previous studies. ML algorithms are designed to automatically learn patterns and insights from data without being explicitly programmed. They use statistical techniques and mathematical models to make predictions or decisions based on the patterns they discover in the data. Without explicit knowledge of the physical consequences of each factor, ML thermal comfort models can discover relationships between occupants' thermal feedback and the influencing variables [5].

Investigations on building comfort have been carried out and suggest that these techniques help to improve the performance of existing comfort models [6]. For example, the work conducted by Yang et al. [7] involved field studies in winter conditions in offices with a sample of 12 subjects. Researchers collected several variables of interest along with participants' thermal sensation, thermal preferences, and thermal satisfaction. They used the ASHRAE scale to assess these factors. Furthermore, researchers using machine learning techniques to analyse the collected data achieved a high level of accuracy.

In thermal comfort studies, it is important to carefully select the most relevant variables to avoid data overload and ensure meaningful insights. Traditionally, comfort studies include many variables, and in some cases, it is possible to find an excessive number of them. Based on that, Ma et al. [8] present in their study the frequency of input variables for the investigation of IAQ and thermal comfort and energy efficiency. Another recent study, by Mamani et al. [9], presents a systematic review of the variables that affect thermal comfort and the instruments used to measure it. However, it is important to strike a balance between the number of variables considered and the overall improvement gained from their inclusion.

This field of investigation requires a standardisation process [10], as exposed below, which allows to identify a minimum and necessary set of variables to carry out a thermal comfort study or for the data collection development proposal, in which a wide range of objective variables or parameters on the indoor and outdoor thermal environment of the building and on personal factors are analysed for the development of models and their implementation.

The main objective of this study is to analyse the variables that influence indoor thermal sensation and to identify the key factors to obtain significant results in future field studies assessing thermal comfort.

To achieve this objective, the following research questions will be investigated.

- What variables have the most significant influence on the indoor thermal sensation in buildings? A comparison between the variables identified in a field study carried out in Seville and the ASHRAE database will provide valuable information on the most important variables that affect thermal sensation. Moreover, variable analysis would contribute to model building and would lead to the optimisation of their performance and gaining insight into the underlying data characteristics for more robust and reliable results. Usually, the datasets contain a large number of entries, but not all of them are relevant or useful for indoor thermal comfort prediction, so analysing the significance of each variable would allow one to identify the most relevant features to improve the efficiency of comfort models.
- Can machine learning algorithms effectively identify the most important variables and contribute to learning and detecting occupants' thermal sensation? Which technique gives the best results? The study aims to investigate the usefulness of machine learning algorithms in the context of thermal comfort and their contribution to the design and

operation of building conditioning systems. Different machine learning techniques will be evaluated to determine which one provides the most accurate and reliable results.

The novelty of this study is based on these two objectives. Although there are references in the literature that focus on analysing indoor and outdoor variables related to the thermal sensation of the occupants of a building and others centre on machine learning algorithms to predict a certain feature, most of them analyse such aspects independently, without considering their interaction. For that, the present paper proposes a novel systematic methodology to explore the importance of variables in the estimation of indoor thermal comfort but by considering different machine learning techniques. In addition, the manuscript analyses the data from a field study carried out in a mixed-mode environment, not usually considered in the literature, and makes a comparison in terms of key variables regarding thermal comfort with other databases previously published (ASHRAE). Researchers in related fields such as building design, climate analysis, or predictive modelling could refer to this study to gain insight into the variables that influence thermal sensation votes. The evaluation of algorithm efficiency also guides the selection of appropriate modelling techniques in future studies.

The present article aims to fulfil the previous aims, proposing an approach to discuss and analyse the relevance of variables with respect to indoor thermal comfort based on two case studies and it could serve as a valuable reference for other studies in this field. For that, Section 2 presents the background and context, and Section 3 discusses the methodology proposed. Section 4 presents the first case study, located in the city of Seville, and analyses the data collection, the importance of the variables, the initial testing of the algorithms, and the selection of hyperparameters. Section 5 presents the results and discussion based on the application of different variables common to the two cases of study considered. Finally, Section 6 presents some conclusions.

2. Background and Context

This section provides relevant background information and situates the current work within the context of previous studies.

The study by Zahra et al. [5] highlights that the growing application of machine learning investigations is driven by its benefits due to its ability to deal with complex problems. Such a study analyses the applications, parameters, methods, performance, and challenges of ML, and highlighted that the results show that 62% of the reviewed studies focused on the development of group-based comfort models. However, Xie et al. [9] conducted a review of thermal comfort studies centred on the occupant [9] with more studies geared toward personal comfort than group comfort. Technologies used to detect inputs in personalised thermal comfort models can indeed be used in the future to predict comfort for individuals or groups. Some smart technologies based on connected sensors, recorders, and wearable devices are reviewed in Čulić et al. [11,12]. For example, an in-office study of 6 days with two subjects was carried out using an infrared camera to analyse thermal comfort in Shanghai [13].

Nevertheless, as [14] indicates, it is necessary to address certain aspects such as the choice between the different ML algorithms, the selection or non-uniformity in the structure of the data, and the considerations that could be included in the selection of the study variables.

The study carried out by [6] compared the performance of different ML algorithms and suggested that factors such as building type, building operating mode, and weather conditions were not among the main factors in predicting thermal comfort. However, the authors promote the study and analysis of these factors, underlying the fact that the thermal perception of the occupants requires further investigation.

Table 1 shows investigations carried out during the last few years regarding thermal comfort, in which climate chambers (CCs) and residential (R), office (O), or university (U) buildings are considered. In studies where the building type is not specified or it can be mixed, it is considered as non-specific (NS).

In refs. [15–18], the ASHRAE databases are used [19], currently a source open to the scientific community. Other studies such as [7,20–25] are based on evidence from their field studies in which they have collected data on environmental conditions and information on subjects from comfort surveys in conditioned (HVAC) office (O) buildings, residential (R) buildings, and universities (U), except for the study [24] where the data are for a naturally ventilated (NV) building.

At present, the application of machine learning-based models allows the inclusion of new variables for the construction of predictive models, so Table 1 also presents a summary of the most recent studies carried out in the area of thermal comfort and in which machine learning techniques are applied.

The table shows the variables used in different studies, including the temperatures inside the building: the indoor air temperature (T_a) and the mean studied at three heights ($\overline{T}_a$) or its measurement at 0.6 m above ground ($T_{a\ 0.6}$), the radiant temperature (T_r), and the mean obtained at three heights ($\overline{T}_r$), the globe temperature (T_g), and the globe temperature measured at 0.6 m. ($T_{g\ 0.6}$).

The temperatures analysed outside the building include: the mean outdoor min/max temperature ($\overline{T}_{out}$), and the outdoor temperature (T_{out}) at specific moments of the day, for example, at 6 a.m. ($T_{out\ 6am}$) or at 3 p.m. ($T_{out\ 3pm}$).

Other physical variables for the space analysed of note include air velocity (v_a) and the mean at three heights ($\overline{v}_a$) or measured at 0.6 m. above ground ($v_{a\ 0.6}$) and relative humidity (RH) and its mean measurement outdoors ($\overline{RH}_{out}$) or its specific measurement at a specific moment in the day, as is the case of the relative humidity obtained at 6 a.m. (RH_{6am}) or at 3 p.m. (RH_{3pm}), as well as CO_2 concentration (CO_2) or partial vapour pressure of water (Pa).

Table 1 also contains user-related variables such as skin temperature (ST), the mean of temperatures obtained on the hands, elbows, shoulders, chest, and head [22], or on the hands (ST_h) in ref. [21], or skin conductance (SC_h) measured in µS, obtained on the hands. In some studies, these temperatures are analysed on clothing, such as the general temperature on clothing (T_{clo}), that is, the average of the temperatures obtained on clothing measured using a thermograph on the hands, elbows, shoulders, and chest [22].

Other variables of note include gender, age, height, weight, clothing insulation (clo), metabolic rate (MET), pulse rate (PR) bpm, blood oxygen level (SpO_2), or blood pressure (BP) measured in mmHg.

In the studies listed in Table 1, the following variables are predominant: T_a, T_r, and RH. However, there are others that also appear frequently such as: v_a, $\overline{T}_a$, and $\overline{T}_r$ [15–17]. Alternative investigations, such as [22], focus only on the temperature of different points in the body and the temperature of clothing as input variables, and the work of [26] incorporates *age* and *gender*.

In addition, in Table 1, it can be observed that the majority of studies list a set of common input variables such as air temperature or relative humidity, and the output variable considered is the thermal sensation vote (TSV), although this vote can be found expressed using a Likert scale of three (TSV 3), five (TSV 5), or seven points (TSV 7). The seven-point thermal sensation scale has been widely used and is considered suitable for describing a one-dimensional relationship between the physical parameters of the interior environment and the subjective thermal sensation; for example, the ASHRAE II database [4] defines the TSV from −3 (cold) to +3 (hot). However, some studies [7,20] group the thermal sensation scales into three points: cold (−3, −2), comfortable (−1 to +1), and warm (+2, +3). Others apply a three-point scale [21] and propose differentiating by gender due not only to variations in physical characteristics but also due to innate physiological differences [27] and therefore recommend the performance of a prediction analysis of the thermal state separated by gender. However, it is usual to find a few extreme values and therefore, in some studies, the use of a five-point scale (TSV 5) is applied or analysed. Here, cold thermal sensations (−3 and −2) and warm thermal sensations (+2 and +3) are grouped, as the scientific community considers that thermal comfort is experienced when certain thermal neutrality is obtained, corresponding to the three middle votes on the thermal sensation scale (−1 and +1). Nevertheless, studies on free positioning scales are also being carried out at present [28].

The studies analysed in Table 1 apply one or more machine learning techniques for classifications with supervised learning [29]: *K nearest neighbour* (KNC), *Naive Bayes* (NB), *support vector machine* (SVC), *Gaussian process classifier* (GPC), *decision tree* (DT), or nonlinear methods, together with linear methods such as *logistic regression* (LR). In addition, some authors [24] apply dimension reduction techniques, such as *linear discriminant analysis* (LDA) or ensemble learning techniques such as *Random Forest* (RF) [30], with either one of the techniques applying a *Neural Network* (NN).

Evaluating the techniques applied, ref. [16] studied RF, SVC, and KNC, obtaining accuracy values of 70.2%, 57.4%, and 67.7%, respectively. In ref. [17] recall values of 48.7% were reached for RF, 48.7% for SVC, and 49.3% for KNC. In ref. [18], accuracy indicators of 50.51% were obtained for SVC, 66% for DT, 43.3% for NB, 80.41% for KNC, 55.67% for NN, and 76.3% for BT. In ref. [22], the accuracy was 80% for KNC, and in ref. [7], an accuracy of 90.3% was reached with DT, 92.9% with KNC, and 93% with SVC. Furthermore, the study by [23] presented a similar result for this last technique, up to 89.8%. In ref. [15], an accuracy of 67.1% was reached using RF for a seven-point TSV (thermal sensation vote), 72.7% for a five-point TSV, and 98.6% for a three-point TSV. The results of this study demonstrate that aggregation improves predictive capacity, but accuracy decreases as the classification size is reduced. Additionally, some studies [7,23,25] have applied ML techniques, focusing on the personalised or individual comfort of the user, together with the analysis of the user average.

Studies have been carried out in user datasets with a minimum size of 6 subjects up to a maximum of 34. The majority of them are located in Asia, in countries with predominantly subtropical and humid continental climates, such as China [7,23–25], Singapore [21,26], and Malaysia [18].

The Mediterranean climate is considered in the studies carried out by [16,17,20], and ref. [6] analysed a warm Mediterranean climate in Peshawar (Pakistan).

Other studies, such as [22], examine data from the ASHRAE database without selecting a specific climate, using locations in Europe, Asia, Australia, and North America. Altan and Ozarisoy [31] showed the conventional methods of design applied by previous studies on thermal comfort, based on worldwide studies on thermal comfort assessment, with statistical methods and sample size, some of which are fully or partially used in the works in Table 1.

The studies in Table 1 focus on periods of a maximum of three months corresponding to the winter and/or summer seasons. Only in ref. [23] did the period exceed one year.

The previous works include datasets of between 485 and 49,966 entries, in which different ML techniques have been applied. Regarding the choice of training, a distribution of 80–90% of training data is generally considered, and 20–10% of test data, depending on the study and the size of the data sample, except for the work of [23], which divides the data into a 50% proportion. Except [15,21], all the references present their results by applying cross-validation [32] as a technique for assessing the results and ensuring that they are independent of the training and test data distribution, with the aim of finding the model that offers the best prediction [17]. The majority of the studies use five divisions (80% of the data for training) and calculate an average of the results obtained with the test data.

After analysing the former studies, the need to deeply investigate new cases is observed, in which the application of ML techniques to other studies is presented, not only for the application in a dataset itself, but also due to the need to consider other countries and climates.

Table 1. Comfort model and automation studies summary.

Ref.	Year	Building, System	Variables		Algorithm(s)—Metrics (%)												
			Input	Output	Accuracy (A) Precision (P) Recall (R) F1-Score (F)	SVC	DT	KNC	RF	NN	LR	LDA	GPC	NB	BT	Train—Test	Size
[23]	2016	CC, HVAC	T_a, T_r, RH, v_a, clo, MET	TSV 7	A	90										50–50%	1199
[26]	2017	O, HVAC R, NV	T_a, T_r, RH, v_a, clo, MET, T_{out}, age, $gender$	TSV 3	A (HVAC) A (NV)	81 73	75 65	75 62		85 72	80 73	77 64				70–30%	812
[21]	2018	O, HVAC	T_a, RH, v_a, T_g, ST, ST_h, SC_h, PR, SpO_2, BP	TSV 3	A (male) A (female)				93 94							80–20%	700
[22]	2019	O, HVAC	ST, T_{clo}	TSV 3	A P R	76 82 76		80 86 80	80 80 80				80 85 80			80–20%	648
[17]	2019	O, HVAC	$\overline{T}_a$, $\overline{T}_r$, RH, $\overline{T}_{out}$, RH_{out}, clo, MET, $\overline{v}_a$	TSV 6	R	49		49	49							90–10%	5576
[20]	2020	O, HVAC	T_a, RH, v_a, T_g, ST	TSV 2 TSV 3	A A	77 75	76 72	80 82	78 77							80–20%	1275
[25]	2021	U, HVAC	T_a, T_r, RH, v_a, clo, MET	PMV	A			88								80–20%	8000
[16]	2021	O, HVAC	$\overline{T}_a$, $\overline{T}_r$, RH, $\overline{v}_a$, clo, MET, Pa, $T_{a\ 0.6}$, $T_{g\ 0.6}$, $v_{a\ 0.6}$	TSV 7	A	57		68	70							90–10%	556
[15]	2021	NS, HVAC	$\overline{T}_a$, $\overline{T}_r$, RH, $\overline{v}_a$, clo, MET, $T_{out\ 6am}$, RH_{6am}, $T_{out\ 3pm}$, RH_{3pm}	TSV 7 TSV 5 TSV 3	A A A				67 73 99							80–20%	7222
[18]	2021	U, HVAC	T_a, T_r, RH, v_a, clo, MET	TSV 5	A	51	66	80		56				43	76	80–10% val.10%	485
[7]	2022	CC, HVAC	T_a, RH, clo, MET, T_g, ST	TSV 3	A	93	90	93								80–20%	6613
[24]	2022	U, NV	T_a, T_r, RH, clo, CO_2, age, $height$, $weight$	TSV 7 TSV 3 TSV 2	A A A	37 59 54	32 49 64	31 48 59			39 56 71	40 55 71		32 48 73		90–10%	49,966
*	2023	O, HVAC	T_a, RH, T_g, T_r, CO_2, T_{out}, $T_{out,\,avg}$, $T_{out,\,max}$, $T_{out,\,min}$, $T_{rm\ 0.8}$, RH_{out}, $RH_{out,\,avg}$, $RH_{rm\ 0.8}$, $gender$, $height$, $weight$, age, clo, $activity$, $emotional$, $food$	TSV 5	A P R F	55 52 55 52	57 55 57 54	55 50 55 49	61 59 61 59					51 42 51 40		80–20%	3352

* This research, with a database of 21 variables from the University of Seville (US extended).

3. Methodology

The methodology used in this study on thermal comfort contains the four stages generally applied when ML is considered: identification, testing, selection of the model, and application [5]. It also includes data collection, preparation, division, construction, and validation of the model.

Previously, a methodological framework could be used in thermal comfort studies, according to the study by Ozarisoy and Altan [33], since in order to develop tailor-made approaches to support the development of thermal comfort, it is essential to conduct a preliminary survey and analysis of field instruments. These preliminary steps involve thermal comfort surveys, the formation of clusters, and the implementation of statistical analysis techniques to gather relevant data and gain insight into the factors influencing thermal comfort. These preliminary steps allow for the collection of relevant data, the identification of the preferences of the occupants through various techniques, and the use of statistical analysis to understand the relationships between variables.

After this methodological framework in field studies, Figure 1 presents the procedure used in this paper in the form of a flow diagram to analyse the variables required in a comfort study in greater detail to obtain good performance from the prediction techniques.

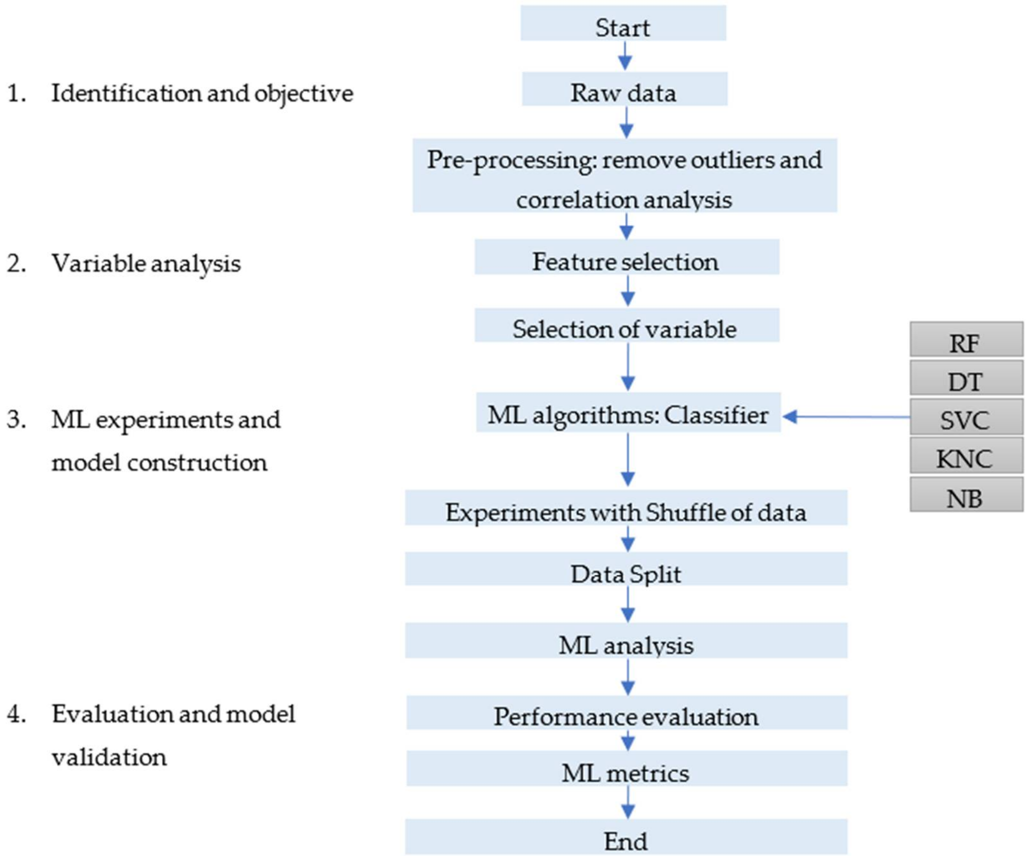

Figure 1. ML process in thermal comfort studies.

Figure 1 presents the procedure used in this paper in the form of a flow chart to analyse in detail the variables required in a comfort study to achieve good prediction performance.

The first stage of the process is the identification of the objectives and the collection of data. The objective of this study is to determine the thermal comfort of a space, based on user responses from a set of input variables. Therefore, a set of sufficient values associated with a set of variables must be collected that allows the thermal response of the user to be detected (raw data). This process, before the application of the different machine learning techniques, involves the collection of data regarding outdoor weather conditions from weather stations, environmental factors from sensors inside the building, and other variables linked to the user of a building, including gender, age, etc., as well as their thermal comfort level obtained from surveys. This unprocessed dataset requires preparation and analysis for processing, so the collected data (database) may require cleaning or filtering. Preprocessing is usually conducted, which eliminates outliers due to sensor errors or loss of input data and performs a correlation analysis.

The second phase focuses on the analysis of variables relevant to indoor thermal comfort. Section 3 develops such an analysis, as it is not common for all the variables in the dataset to be equally useful for defining a model: if redundant variables are added, the generalisation capacity of the model may decrease, and the general precision of a classifier may also be reduced. An analysis of the relevance of the variables follows below, so the performance of the models as variables are added in order of importance is studied, as adding more variables to a model increases the general complexity of the model. Therefore, an analysis is made in which the process of selecting the variables of interest is presented. This implies the elimination of those variables whose information is very similar and likely to be redundant, as they are highly correlated (*remove highly correlated variables*). This is followed by a selection of the most representative variables or characteristics for the problem under study from those collected in the database (*feature selection*).

Although there are different ML techniques, ref. [34] provides a comprehensive overview of machine learning algorithms and their real-world applications, and in refs. [35,36], the authors present a concise explanation of these techniques.

In the third phase, different ML classification algorithms were compared. To demonstrate the quality of the techniques, the data are shuffled, and several tests are run in which the dataset is split for training and testing.

Every machine learning technique has a set of hyperparameters that must be adjusted since the selection of these hyperparameters has a direct impact on the performance of the model [37]. A search space must first be defined for the selection of these hyperparameters, formed by a set of tuned ranges. The combinations of hyperparameters in the search space are then detected to finally select the combination with the best performance. In particular, the grid search (GS) method permits an exhaustive exploration to be carried out or a brute force method which tests all the combinations of hyperparameters [38] until finding the optimum configuration. Nevertheless, if the implementation time or resources are limited, an alternative may be a random search (RS) [39] or Bayesian optimization [40].

The fourth stage consists of evaluating metrics that lead to the identification of the best performing model (the performance metrics will be identified in Section 4.1.3).

4. Case Studies

This study aims to discuss and analyze the variables regarding indoor thermal comfort of fieldwork carried out in Seville, Spain, together with an analysis and comparison with the Mediterranean climate dataset developed from the ASHRAE database. The Seville study is described, including the description of the variables, the preparation of the data, and results relating to the testing and selection of variables and hyperparameters. Lastly, the Mediterranean climate case study based on the ASHRAE database is also examined.

4.1. University of Seville

The University of Seville (US) study is based on data collected in a field study described in ref. [41], considering different areas in university buildings dedicated to office work with 54 occupants mainly aged between 20 and 40 years of age.

The field study was carried out in the city of Seville (Spain), where the climate is classified as a temperate climate, ‘Csa’ (temperate, with dry hot summers, hot summer Mediterranean), according to the Köppen–Geiger classification [42]. In particular, the climate of the city of Seville is characterised by mild winters and very hot summers, with mean temperatures above 10 °C in winter and 30 °C in summer.

In this study, the winter and summer comfort data are analysed together because combining all data from one year provides a more comprehensive understanding of overall thermal comfort in a given environment. By considering data from both seasons, the results may reveal patterns, trends, or correlations that may not be apparent when each season is analysed separately. This approach could be applied to systems, without the need to identify the seasons of the year and to identify variables to optimise occupant comfort in different seasons. According to Arakawa et al. [43], analysing a single season can be limiting when trying to capture the full range of thermal sensations and preferences.

Eleven office rooms in three blocks, two of which were mixed-mode buildings, that combined natural ventilation and support conditioning systems controlled by the user, while the third building was completely conditioned, were analyzed. The study included the collection of indoor and outdoor variables, as well as the thermal sensation of the occupants through surveys.

4.1.1. Data Collection

The results of 3352 surveys were used. From these, the following variables are known and form the set of input variables for the ML algorithms and relate to:

- The occupants: *gender*, *height*, *weight*, and *age*.
- The adaptive actions: clothing (*clo*), *activity*, *emotional* state, and *food* consumption.
- The inside of the building: indoor air temperature (T_a), indoor relative humidity (RH), globe temperature (T_g), radiant temperature (T_r), and indoor CO_2 level (CO_2).
- The outside of the building: outdoor air temperature (T_{out}), outdoor air temperature based on ISD (Integrated Surface Dataset) data ($T_{out,\ avg}$) together with the maximum value of the day ($T_{out,\ max}$) and the minimum value of the day ($T_{out,\ min}$) and also the running mean outdoor temperature with a weighted factor of 0.8 ($T_{rm\ 0.8}$). Furthermore, outdoor relative humidity (RH_{out}), outdoor relative humidity based on ISD data ($RH_{out,\ avg}$), and running mean relative humidity with a weighted factor of 0.8 ($RH_{rm\ 0.8}$) were also collected.

The variables $T_{rm\ 0.8}$ and $RH_{rm\ 0.8}$ are calculated based on a running mean [44], which refers to a method of calculating an average or mean value by using a sliding window or moving average over a series from the mean of the instant outdoor value for each day of the study (T_{od}). The weighted factor, α, can be seen as a smoothing constant ($0 < \alpha < 1$) that quantitatively reflects the rate at which the effect of any past temperature or humidity declines. The higher the value, the greater the effect of the past temperature or humidity. The value 0.8 is the usual value in most studies and is considered the best metric according to the study by [45].

In summary, a total of 21 variables were collected, which will be analysed in depth below regarding their importance in indoor thermal comfort.

4.1.2. Data Preparation

Thermal sensation, determined by the thermal sensation vote, serves as the output variable in the present study and is used in the learning process to assess indoor thermal comfort. In this study, a thermal sensation vote is applied on a five-point Likert scale for learning by classification. This is a multiclass output consisting of the values -2, -1, 0, 1, and 2.

An exploratory analysis of the data from the case study was performed, with respect to the relevance of the input variables to predict the thermal sensation vote. Figure 2 shows the importance of the 21 study variables (US extended).

The factor of the most importance is T_a, followed by T_r, T_g, CO_2, *age*, *RH*, $T_{rm\ 0.8}$, T_{out}, $RH_{rm\ 0.8}$, RH_{out}, and *clo* (US minimum).

Furthermore, a correlation analysis of the variables was performed, although multicollinearity did not create problems in prediction capacity but in interpretability, since this does not reduce the predictive capacity of the model and only affects calculations related to individual predictors.

Figure 3 shows the Pearson correlation analysis of the variables, where the correlation was found to be below the limit of ± 0.8. Therefore, the variables have a high degree of linear independence; therefore, the predictor variables analysed were independent. However, some features of interest are moderately dependent; for example, *clo* is moderately correlated with outdoor variables of temperature and humidity, as has been shown previously in other studies in the same climate zone [46]. It was observed that the variables gender, height, and weight are moderately correlated, and among the physical variables, there is a moderate correlation between the outdoor humidity and temperature variables.

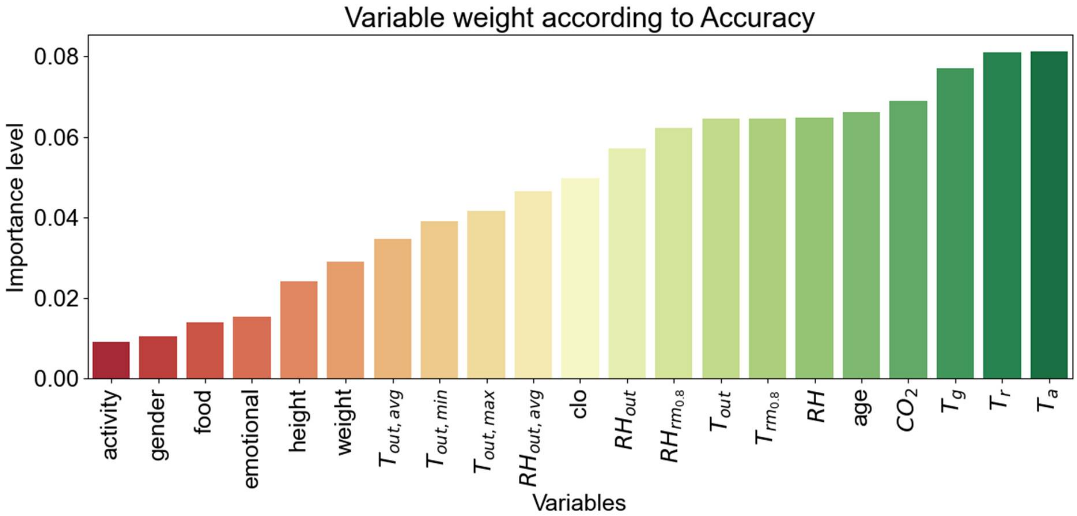

Figure 2. Level of importance of each variable.

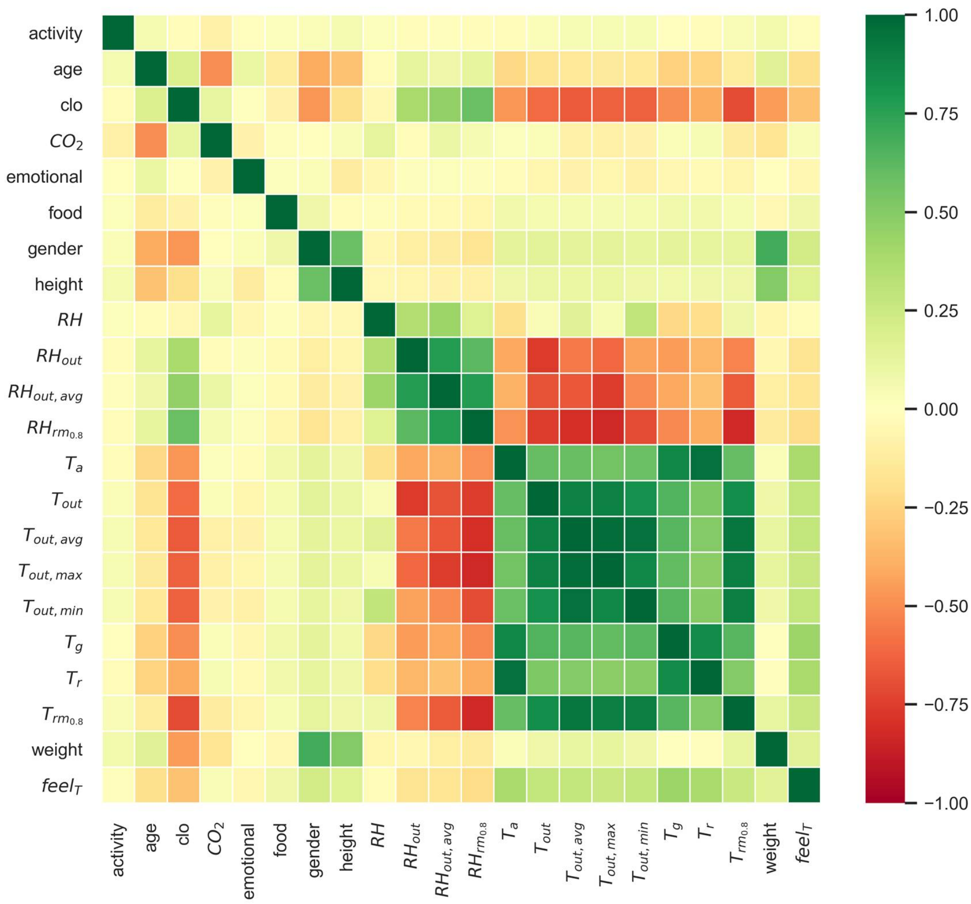

Figure 3. Pearson correlation.

Based on technical considerations of the algorithms, the only reason to exclude these features from the study would be based on storage and resolution speed, as the data contribute to the multi-class prediction and their data quality is sufficient. Variable selection involves a trade-off between computational efficiency and the inclusion of potentially valuable variables. In cases where storage and computational constraints exist, it may be necessary to prioritise variables that have a higher perceived value or relevance to the indoor thermal comfort experience. However, it is important to ensure that the remaining variables provide a complete representation of the problem domain and capture the information necessary for accurate predictions.

Based on the field study or the implementation of these variables in the system, it is important to know the most relevant variables since it may be necessary to exclude certain variables from the study to manage costs effectively (the cost of installing sensors and collecting, and archiving data).

4.1.3. Testing the Algorithms and Results

The analysis of the behaviour of the techniques permits the identification of which ML models offer the best performance. For that, testing based on the random selection of the dataset was performed to achieve a sample size that contains a diverse range of situations, regardless of the data, time, or study period (winter or summer). A thousand random selections of the data were constructed, to which cross-validation was applied, with a K-fold of 5. This implies using 80% of the data for learning and 20% for testing. Thus, an extensive comparative sample of the following classification algorithms will be presented: random forest [47], decision tree [48], support vector machine [49], K-nearest neighbours [46], and Naive bayes [50].

The techniques with the selected hyperparameters were analysed and evaluated on the basis of a set of metrics that assess the performance of the models.

Accuracy (Equation (1), Table 2) provides the exactitude by representing the proportion of both positive and negative cases that the technique has correctly classified, that is, those that were predicted to respond to the comfort level and those that really do. The precision metric (Equation (2), Table 2) detects the ability to predict the correct classification of a situation. As the objective is to detect user comfort, this metric provides information about the precision of the classification response, given that TSV consists of n points. Recall (Equation (3), Table 2) represents the exhaustivity or proportion of true predicted positives of the total of true positives, that is, the number of user responses that would be correctly predicted based on the ASHRAE scale. Lastly, the F1-score (Equation (4), Table 2) is the metric that reflects the combination of precision and recall.

Table 2. Metrics.

Metric	Basic Classification	Multi-Class Classification		Equations
	i-Class	Macro Avg	Weighted Avg	
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP+TN}{TP+TN+FP+FN}$	(1)
Precision	$\frac{TP}{TP+FP}$	$\frac{Prec_1+Prec_2+\ldots+Prec_n}{n}$	$Prec_1 \times Prop_1 + Prec_2 \times Prop_2 + \ldots + Prec_n \times Prop_n$	(2)
Recall	$\frac{TP}{TP+FN}$	$\frac{Recall_1+Recall_2+\ldots+Recall_n}{n}$	$Recall_1 \times Prop_1 + Recall_2 \times Prop_2 + \ldots + Recall_n \times Prop_n$	(3)
F1-score	$2 \cdot \frac{Prec \cdot Recall}{Prec+Recall}$	$2 \cdot \frac{Prec \cdot Recall}{Prec+Recall}$	$2 \cdot \frac{Prec \cdot Recall}{Prec+Recall}$	(4)

As this is a problem of multiclass classification, the metric is calculated for each class and the weighted average is considered (Table 2) according to the number of samples per class. Although each averaging method has its advantages, a common consideration when selecting the right method is the imbalance of classes: if the classes have different sample numbers, it may be more accurate to use a macros average (Table 2), where the minority classes receive the same weight as the majority classes.

The results of the application of the methodology lead to the selection of the ML technique with the best performance from among those that have been trained and run in the previous stage. Moreover, the running time and computational costs of the models must also be considered [5]. Such technique selected will be the preferred model for application in studies on thermal comfort carried out in buildings, which seek to offer occupants thermal comfort using a conditioning system based on the comfort learning expressed by the users.

The results shown in Figure 4 demonstrate that, among the whole algorithms considered, the technique offering the best performance to the dataset with 21 variables (US extended) is random forest, although it can be observed that the metrics that apply the weighted average offer a better perspective of the result.

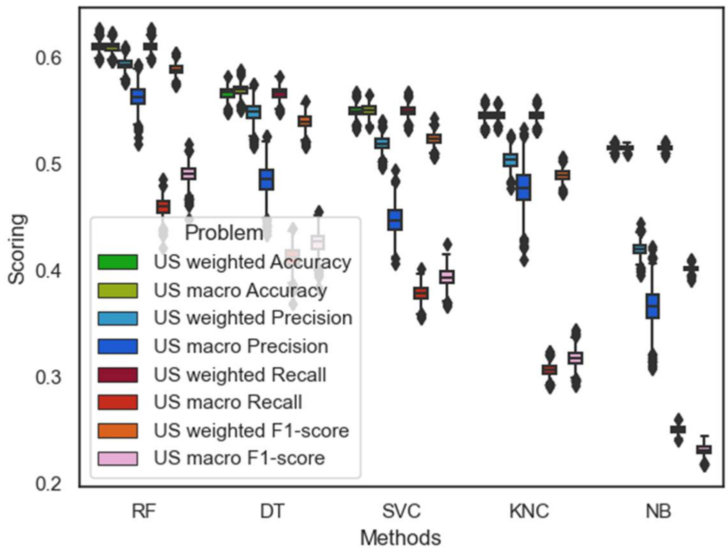

Figure 4. University of Seville case study extended with 21 variables.

4.1.4. Selection of Hyperparameters and Effect of the Number of Study Variables

To improve the performance of the machine learning models, their hyperparameters are usually tuned. This study has chosen to use GridsearchCV, whose operation involves running the model using cross-validation to find the best combination of hyperparameters. Once the best combination of parameters in the extended University of Seville database has been obtained, Table 3 shows the performance of the ML algorithms for the case study extended with 21 variables (US extended) and 8 variables (US minimum).

Table 3. Tuning of hyperparameters with the University of Seville database.

Parameters and Hyperparameters in the US (Extended)			Scoring	US Extended		US Minimum		Improvement over the Case with 8 var. (%)
Methods	Analysis Range	Optimum		Mean	std	Mean	std	
RF	Number of trees (1; three hundred)	181	Accuracy	0.611	0.004	0.587	0.004	4%
			Precision	0.594	0.005	0.560	0.006	6%
	Maximum depth of the tree (1; 20)	13	Recall	0.611	0.004	0.587	0.004	4%
			F1-score	0.589	0.005	0.555	0.004	6%
DT	Maximum depth of the tree (1; 20)	8	Accuracy	0.567	0.005	0.540	0.007	5%
			Precision	0.549	0.009	0.516	0.008	6%
	Minimum number of samples required to be at a leaf node (5;100)	70	Recall	0.567	0.005	0.540	0.007	5%
			F1-score	0.540	0.007	0.521	0.007	4%

Table 3. *Cont.*

<table>
<tr><th rowspan="2">Methods</th><th colspan="2">Parameters and Hyperparameters in the US (Extended)</th><th rowspan="2">Scoring</th><th colspan="2">US Extended</th><th colspan="2">US Minimum</th><th rowspan="2">Improvement over the Case with 8 var. (%)</th></tr>
<tr><th>Analysis Range</th><th>Optimum</th><th>Mean</th><th>std</th><th>Mean</th><th>std</th></tr>
<tr><td rowspan="4">SVC</td><td rowspan="2">Regularization parameter (0.1; 1000)</td><td rowspan="2">100</td><td>Accuracy</td><td>0.550</td><td>0.005</td><td>0.538</td><td>0.004</td><td>2%</td></tr>
<tr><td>Precision</td><td>0.520</td><td>0.006</td><td>0.495</td><td>0.009</td><td>5%</td></tr>
<tr><td rowspan="2">Kernel coefficient, gamma (0.0001; 1)</td><td rowspan="2">0.0001</td><td>Recall</td><td>0.550</td><td>0.005</td><td>0.538</td><td>0.004</td><td>2%</td></tr>
<tr><td>F1-score</td><td>0.524</td><td>0.005</td><td>0.482</td><td>0.004</td><td>9%</td></tr>
<tr><td rowspan="4">KNC</td><td>Number of neighbours (1; 50)</td><td>16</td><td>Accuracy</td><td>0.546</td><td>0.004</td><td>0.525</td><td>0.004</td><td>4%</td></tr>
<tr><td>Leaf size (1; 50)</td><td>1</td><td>Precision</td><td>0.504</td><td>0.008</td><td>0.471</td><td>0.008</td><td>7%</td></tr>
<tr><td>Power parameter for the Minkowski metric (1; 2)</td><td>1</td><td>Recall</td><td>0.546</td><td>0.004</td><td>0.525</td><td>0.004</td><td>4%</td></tr>
<tr><td>metrics (‘Minkowski’; ‘Euclidean’; ‘Manhattan’) weights (‘uniform’; ‘distance’)</td><td>Minkowski distance</td><td>F1-score</td><td>0.490</td><td>0.005</td><td>0.445</td><td>0.004</td><td>10%</td></tr>
<tr><td rowspan="4">NB</td><td rowspan="4">Portion of the largest variance (0; 10–15)</td><td rowspan="4">0.00187</td><td>Accuracy</td><td>0.516</td><td>0.002</td><td>0.518</td><td>0.002</td><td>−1%</td></tr>
<tr><td>Precision</td><td>0.420</td><td>0.006</td><td>0.456</td><td>0.019</td><td>−8%</td></tr>
<tr><td>Recall</td><td>0.516</td><td>0.002</td><td>0.518</td><td>0.002</td><td>−1%</td></tr>
<tr><td>F1-score</td><td>0.402</td><td>0.003</td><td>0.411</td><td>0.002</td><td>−2%</td></tr>
</table>

For the *random forest* algorithm, 181 trees were selected with a depth of 13. A number of trees between 1 and 300 were analysed, with a depth of between 1 and 20, extending the study range of [51].

For the *decision tree* algorithm, a depth of 8 trees and 70 samples was selected. For that, a depth of between 1 and 20 was analysed, together with a minimum of 5 to 100 samples, extending the study range proposed by [51].

For the *support vector classifier* hyperparameters, a cost of 100 and a gamma of 0.0001 were selected considering the intervals *cost* [0.1 to 1000] and *gamma* [1 to 0.0001], respectively, based on the study by [52].

For the *K-nearest neighbours* technique, the hyperparameters were optimised, obtaining 16 as the number of neighbours (in an interval from 1 to 50, extending the study range of [51]), together with a leaf size of 1 (in the interval 1 to 40 defined by [51]). It is noted that the number of neighbour parameters obtained is close to the proposal of [24]. Moreover, the selected power parameter (p) was 1, for the weights obtained with a distance metric, by applying the *Minkowski* method.

Lastly, ref. [51] proposes the application of a Gaussian distribution for the *naive Bayes* technique, as this is the technique that gave the best performance. This distribution was also applied in this study, as the multinomial distribution does not allow the use of negative variables (below zero temperature). The smoothing parameter was also analysed and a parameter of 1.87×10^{-3} was obtained.

When a study is configured with fine-tuning of the hyperparameters, each hyperparameter to be adjusted is defined by its type and the range of values for which these parameters have been analysed. Grid search is a technique conventionally applied by checking all the possible combinations of parameters evaluated. It is considered a method that is easy and simple to implement [39], although the method requires a lot of time when the number of parameters is increased.

One of the objectives of this study is to determine how many variables are necessary for a comfort study to obtain an accurate estimation of it. Considering that the application of random forest produces good results, the evolution of the number of variables was analysed in the order of importance suggested in Figure 2.

In all the cases, 1000 random sets were tested, with optimum hyperparameter values. It was observed that above eight variables (US minimum), with the highest level of importance, the improvement tended to stabilise, as can be observed in Figure 5. Therefore, these eight variables can be considered as fundamental in a comfort study (minimum). As can be observed in Table 3, the rest of the variables can provide an improvement of around 5%.

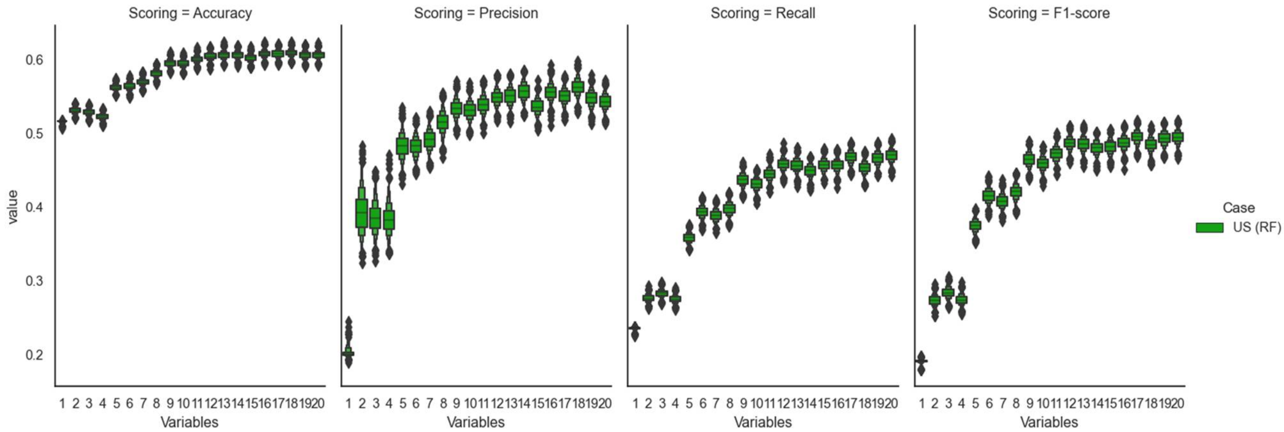

Figure 5. Evolution of the indicators with an increasing number of variables.

To highlight this variation in the different techniques, Figure 6 shows the results in two situations: selection of all input variables (US extended) and selection of the eight most important (T_a, T_r, T_g, CO_2, *age*, *RH*, $T_{rm\ 0.8}$, and T_{out}), that is, the 'US minimum'.

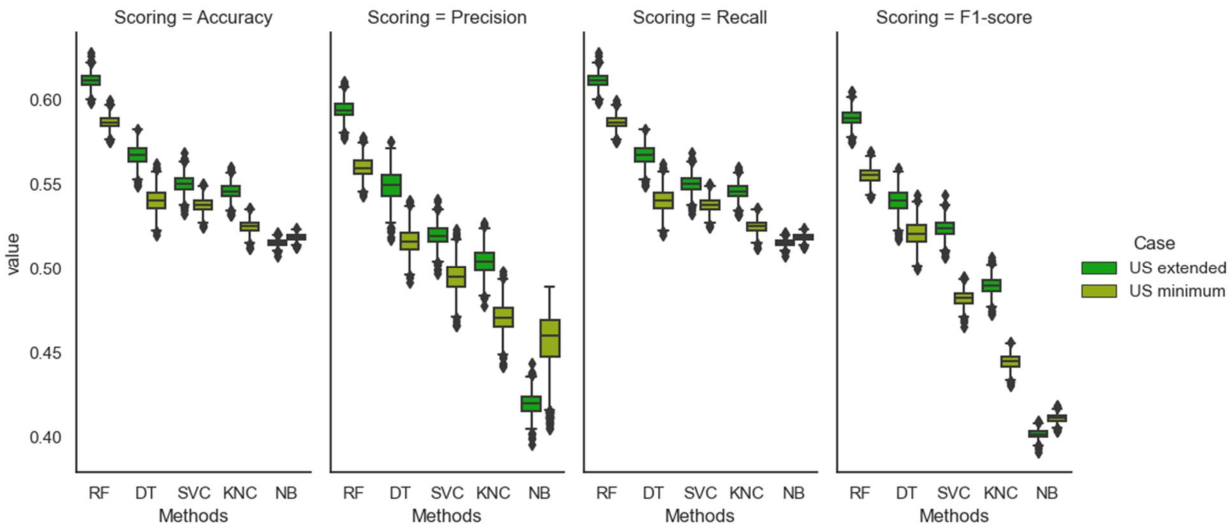

Figure 6. Comparison of metrics between the different machine learning techniques for the database of the University of Seville with 21 variables (US extended) and the 8 main variables (US minimum).

In Figure 6, it can be seen that the metrics improve in all the models when all the variables are used, except NB.

4.2. ASHRAE

In order to explore the relevance of variables related to indoor thermal comfort, we propose to analyse and compare the case study described above with a significant dataset similar to the case study, formed by a set of entries in the ASHRAE II database [4] relevant to the Mediterranean climate and office buildings. For that, a dataset was defined, which permits the methodology to be verified and the results to be compared. From the ASHRAE II database (https://datadryad.org/stash/dataset/doi:10.6078/D1F671 (accessed on 19 July 2023)), the Mediterranean climate data were selected according to Köppen–Geiger [53], involving locations belonging to 'Csa' and 'Csb' (temperate with dry warm summers, warm summer Mediterranean) that are characterised by temperatures and rainfall in climates similar to the climate zone of Seville, which corresponds to 'Csa' (temperate, with dry hot summers, hot summer Mediterranean). A total of 5569 entries were obtained considering different climate zones: Athens (Greece, July–September 1993 and July–August 1994) [54], Auburn (east coast of the USA, March 1993) [55], Berkeley (west coast of the USA, September–October 2002 and February–March 2003) [56], San Román (west coast of the USA, April 1991, December 1992, and July 1993) [55], and Oporto (Portugal, September 1998 to September 1999) [57], defining a dataset of similar size to the study carried out in Seville.

The studies located in Athens, Berkeley, and Oporto were carried out in buildings with natural ventilation (NV). On the other hand, the buildings in the studies in Auburn and San Román had an air conditioning system (AC). The study carried out in Seville took place in mixed-mode (MM) buildings.

Regarding the input variables from the ASHRAE II database, the only one related to the occupants is the gender variable, and the one related to the adaptive actions is the level of clothing (*clo*) and the metabolic rate (*met*). The variables inside the building are the indoor air temperature (T_a) and the radiant temperature. The variables for the building exterior are outdoor air temperature (T_{out}), outdoor air temperature based on ISD data ($T_{out,\,avg}$), beside the maximum value of the day ($T_{out,\,max}$), the minimum value of the day ($T_{out,\,min}$), and the running mean outdoor temperature with a weighted factor of 0.8 ($T_{rm\ 0.8}$) together with the following humidity-related values: outdoor relative humidity based on ISD data ($RH_{out,\,avg}$) and the running mean relative humidity with a weighted factor of 0.8 ($RH_{rm\ 0.8}$).

Table 4 presents the variables from this study of the ASHRAE database for the Mediterranean climate, together with the variables presented in the University of Seville study. The output variable in both studies is the thermal sensation vote based on a Likert scale (multi-class).

Table 4. Variables in both studies.

	Variables	University of Seville	ASHRAE	University of Seville			ASHRAE		
				Min	Avg	Max	Min	Avg	Max
Human variables	*gender*	✓	✓	0	0.72	1	0	0.46	1
	height	✓		158	175	187			
	weight	✓		52	73	120			
	age	✓		26	32	60			
Adaptive actions	*clo*	✓	✓	0.34	0.89	1.63	0.26	0.61	1.64
	activity	✓		−1	0.03	1			
	emotional	✓		−1	0.16	3			
	food	✓		0	0.31	2			
	met		✓				0.8	1.23	3.8
Indoor variables	T_a	✓	✓	13	23	64	17	26	39
	RH	✓		16	4	79			
	T_g	✓		13	23	34			
	T_r	✓	✓	13	23	34	14	26	40
	CO_2	✓		395	735	1907			
Outdoor variables	T_{out}	✓	✓	2	19	42	4	19	28
	$T_{out,\,avg}$	✓	✓	5	17	33	3	1	31
	$T_{out,\,max}$	✓	✓	9	23	44	9	23	36
	$T_{out,\,min}$	✓	✓	0	11	26	−7	14	28
	$T_{rm0.8}$	✓	✓	9	17	31	5	19	29
	RH_{out}	✓		9	60	97			
	$RH_{out,\,avg}$	✓	✓	22	65	96	25	65	95
	$RH_{rm0.8}$	✓	✓	32	65	88	35	65	87

The university study contains more variables, relating both to the occupant and to the adaptive actions or values in the building interior, as can be seen in the table. The only exterior variable that both studies do not share is the variable RH_{out}, from the University of

Seville database, which represents the momentary daily outdoor humidity values, while the variable $RH_{out,\,avg}$ includes mean values for the daily outdoor humidity.

5. Results and Discussion

The different techniques used in both case studies and the influence of the variables will be analysed. Therefore, firstly both databases are compared for the same set of variables. Secondly, an analysis of the typical variables of the University of Seville database is performed, which is compared to the selected ASHRAE database with its typical variables.

5.1. Analysis of the Variables Common to Both Studies

A comparison follows of the results obtained in both datasets by applying the ML techniques considered. For that, datasets formed from the input variables common to the Seville database and the ASHRAE database are trained. These variables are those shown in Table 4: *gender*, *clo*, T_a, T_r, $T_{out,\,avg}$, $T_{out,\,max}$, $T_{out,\,min}$, $T_{rm\ 0.8}$, $RH_{out,\,avg}$, and $RH_{rm\ 0.8}$.

Table 5 presents the results after testing 1000 randomly shuffled sets from both databases (USc and ASHRAEc) to obtain the average test results. The results obtained for the University of Seville database (USc) were confirmed to be slightly better than those for the ASHRAE database (ASHRAEc). Ref. [58] demonstrates that a mixed-mode model (MM), a unified model, is equivalent to an AC or NV when predicting the TSV as these ML techniques have an equivalent performance to the independent AC and NV models.

Table 5. Comparison of the classification models based on scoring, with common variables in the US database (US) and the ASHRAE database (ASHRAE).

		University of Seville (Usc)		ASHRAE (ASHRAEc)	
Methods	**Scoring**	**Mean**	**std**	**Mean**	**std**
RF	Accuracy	0.591	0.004	0.489	0.004
	Precision	0.569	0.006	0.479	0.006
	Recall	0.591	0.004	0.489	0.004
	F1-score	0.561	0.005	0.462	0.005
DT	Accuracy	0.551	0.004	0.476	0.005
	Precision	0.516	0.011	0.450	0.008
	Recall	0.551	0.004	0.476	0.005
	F1-score	0.485	0.005	0.445	0.006
SVC	Accuracy	0.563	0.003	0.489	0.003
	Precision	0.526	0.019	0.484	0.007
	Recall	0.563	0.003	0.489	0.003
	F1-score	0.494	0.003	0.435	0.003
KNC	Accuracy	0.558	0.004	0.473	0.003
	Precision	0.530	0.007	0.451	0.006
	Recall	0.558	0.004	0.473	0.003
	F1-score	0.509	0.005	0.427	0.004
NB	Accuracy	0.505	0.001	0.446	0.001
	Precision	0.290	0.023	0.371	0.014
	Recall	0.505	0.001	0.446	0.001
	F1-score	0.341	0.001	0.374	0.001

Figure 7 shows the range of variability of the performance of each metric using the same variables in each database. It is observed that, in general, the ranges of values corresponding to the USc database are higher than those of the ASHRAEc. For both case studies, random forest is the technique with the best performance.

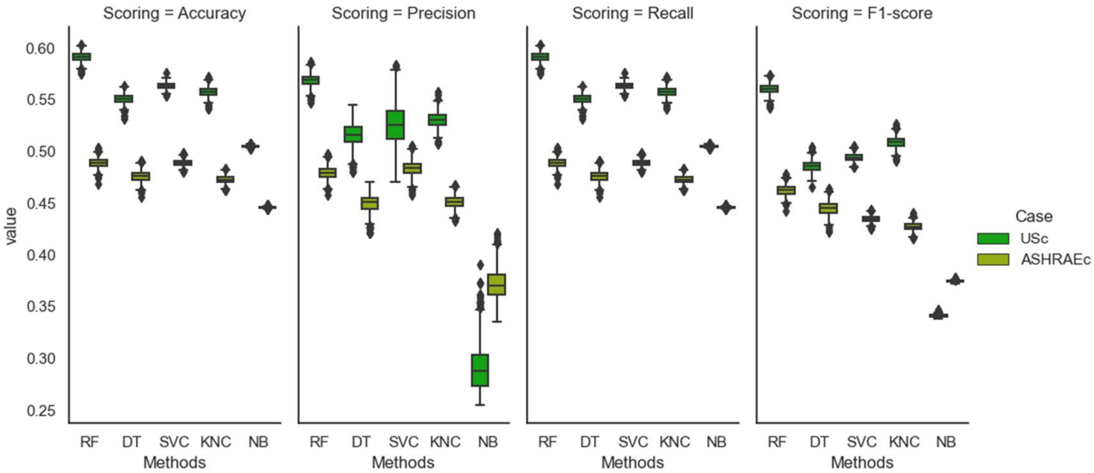

Figure 7. Comparison of the classification models based on scoring, with common variables in the US database (Usc) and the ASHRAE database (ASHRAEc).

Figures 8 and 9 show the importance of each of the 11 USc and ASHRAEc factors in the process used to predict output when the variables common to both studies are considered. As indicated in the study by [59], decision tree algorithms can show the importance of the variables to be trained, although ref. [58] highlights that, compared to defining deep DT with highly dimensioned characteristics, RF constructs a number of smaller trees and suffers less over-adjustment. Thanks to this, RF can develop models better than those developed by DT. In this study, as in refs. [16,51], the importance of input characteristics is calculated using the RF classifier.

Regarding the relevance of variables in indoor thermal comfort, as in the University of Seville study, the most significant are indoor air temperature and radiant temperature, as is the case in the ASHRAE study. From the comparison, it can be seen that outdoor temperature and clothing are two variables with significant influence in both studies. Although, it can be seen that, in the case of the US database, relative humidity has a significant influence.

Running mean humidity and temperature with a weight factor of 0.8 from the US study are significantly more important than those of ASHRAE. On the other hand, gender and the maximum/minimum outdoor temperatures are the variables with least weight in both studies.

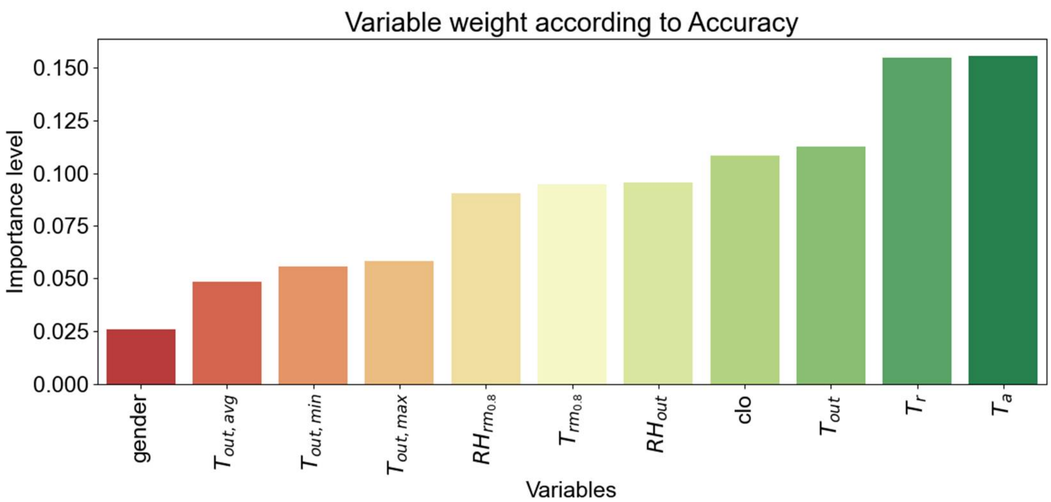

Figure 8. Weight of Usc variables (common variables with the ASHRAEc study) using RF.

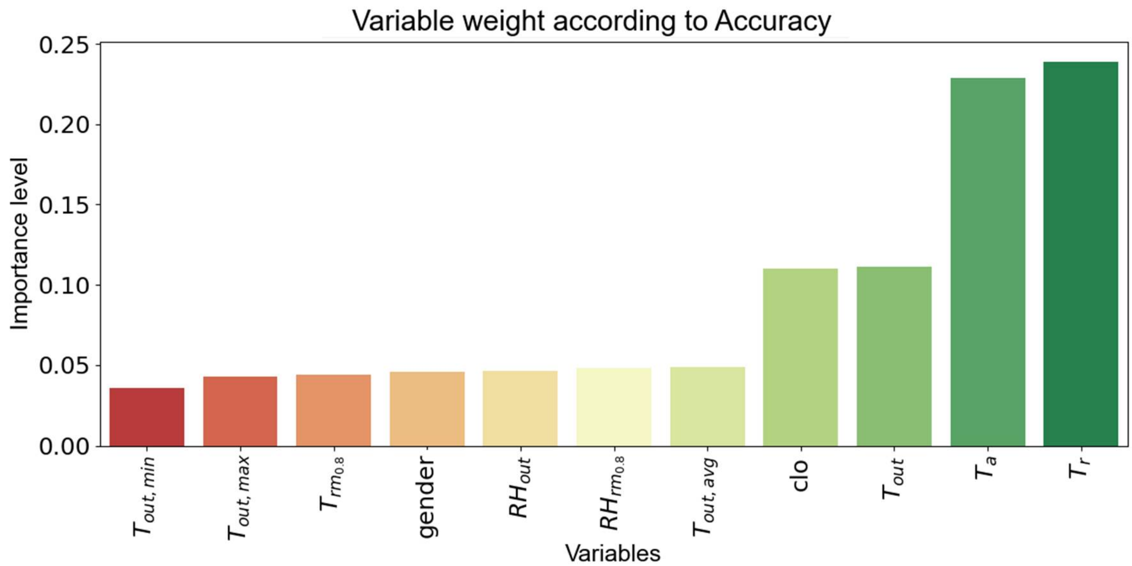

Figure 9. Weight of ASHRAEc variables (common variables with the Usc study) using RF.

5.2. Analysis of the Typical Variables of Each Database

After analysing the results of both models for common variables, the study was extended to include a comparison of the typical variables in each study (Table 4). Similarly, classification metrics are obtained for each ML technique in the ASHRAEo database, and the results are compared with those obtained in the University of Seville database.

Table 6 shows that the extension of the number of variables is favourable in the resolution of all of the models. The results shown in Table 6 correspond to the mean values of accuracy, precision, recall, and F1-score, together with the typical deviations of the values set of each of these metrics.

Better results were obtained with the University of Seville (USo) field study database using typical variables than with the ASHRAEo database. As in the study by [26] for air-conditioned and naturally ventilated buildings, better results are obtained when using classification techniques when the dataset contains Fanger parameters (T_a, T_r, *HR*, v_a, *clo*, and *met*), and additionally, the variables *age* and *gender* and the external ambient parameters also help to improve the result.

Table 6. Comparison of classification models based on scoring, with own variables in the US database (USo) and ASHRAE database (ASHRAEo).

		University of Seville (USo)		Improvement over the USc (%)	ASHRAE (ASHRAEo)		Improvement over the ASHRAEc (%)
Methods	**Scoring**	**Mean**	**std**		**Mean**	**std**	
RF	Accuracy	0.611	0.004	3%	0.498	0.004	2%
	Precision	0.594	0.005	4%	0.494	0.006	3%
	Recall	0.611	0.004	3%	0.498	0.004	2%
	F1-score	0.589	0.005	5%	0.465	0.004	1%
DT	Accuracy	0.567	0.005	3%	0.475	0.005	0%
	Precision	0.549	0.009	6%	0.450	0.009	0%
	Recall	0.567	0.005	3%	0.475	0.005	0%
	F1-score	0.540	0.007	11%	0.443	0.007	−1%
SVC	Accuracy	0.550	0.005	−2%	0.485	0.004	−1%
	Precision	0.520	0.006	−1%	0.475	0.007	−2%
	Recall	0.550	0.005	−2%	0.485	0.004	−1%
	F1-score	0.524	0.005	6%	0.447	0.004	3%

Table 6. *Cont.*

Methods	Scoring	University of Seville (USo) Mean	std	Improvement over the USc (%)	ASHRAE (ASHRAEo) Mean	std	Improvement over the ASHRAEc (%)
KNC	Accuracy	0.546	0.004	−2%	0.474	0.004	0%
	Precision	0.504	0.008	−5%	0.456	0.006	1%
	Recall	0.546	0.004	−2%	0.474	0.004	0%
	F1-score	0.490	0.005	−4%	0.428	0.004	0%
NB	Accuracy	0.516	0.002	2%	0.446	0.001	0%
	Precision	0.420	0.006	45%	0.372	0.014	0%
	Recall	0.516	0.002	2%	0.446	0.001	0%
	F1-score	0.402	0.003	18%	0.374	0.001	0%

Regarding the performance of the ML algorithms considered, the random forest classifier obtains the best results in both databases. From the classification analysis, if only the class corresponding to the neutral situation is analyzed, the method obtains a recall of 0.78, higher than the average. However, lower recall values were obtained for the other classes.

In the literature review, Table 1, similar performances were obtained for the recall metric. Additionally, the study by [58] describes results equivalent to those obtained in Table 1, and as highlighted in the US case study, the value obtained was 61.1% for the University of Seville database with the typical variables.

It can be seen in Table 1, that in all the models analysed in which the typical variables are used, the technique with the highest first quartile is RF, followed by SVC and KNC. The highest value of RF corresponds to an accuracy of 65%, which is 10% higher than the highest value of DT and KNC (≈60%). In the case corresponding to the typical variables of ASHRAE, this difference is even smaller, as the highest accuracy values for these three techniques are around 50% as shown in Table 1. The standard deviations are similar in all studies.

Similar results can also be found in the literature. For example, in the case of [51], the RF and KNC algorithms achieve accuracy values greater than 60%. SVC and DT obtain lower scores, in the range of 56% to 59%, although they have outliers greater than 60%. The NB technique has the worst accuracy, which depends on the distribution used in the model. Using the Gaussian, polynomial, or Bernoulli distribution, performances of 49.8%, 52.5%, and 54.6% were obtained, respectively.

In this study (Figure 10 and Table 6), accuracy scores for RF were obtained that reached maximums of 65% and around 51% for the US and ASHRAE databases, respectively. For DT, maximums of 60% and 50% were reached, respectively. SVC obtained the worst scores, 55% and 37% for each of the databases. KNC obtained 58% and 50%, and NB 57% and 37%, using the Gaussian distribution in this last algorithm.

The behaviour of the ML considered when the number of variables is increased, from the USc and ASHRAEc (common variables) to USo and ASHRAEo (own variables), is an improvement of the metrics (accuracy, precision, recall, and F1-score) in RF and DT up to 3% (Table 6, improvement over the USc and ASHRAEc), and a decrease or not increased in KNC, SVC, and NB (except NB in USo cases, although it is still the worst model).

This is because the RF and DT models are capable of handling a large number of variables, including both relevant and less relevant ones. These algorithms can automatically perform feature selection and identify the most informative variables for prediction. Extending the number of variables in these models can potentially enhance their ability to capture complex relationships and patterns, resulting in improved accuracy, precision, recall, and F1-score. These two models are the ones that worked best for both common (USc and ASHRAEc) and own (USo and ASHREo) cases.

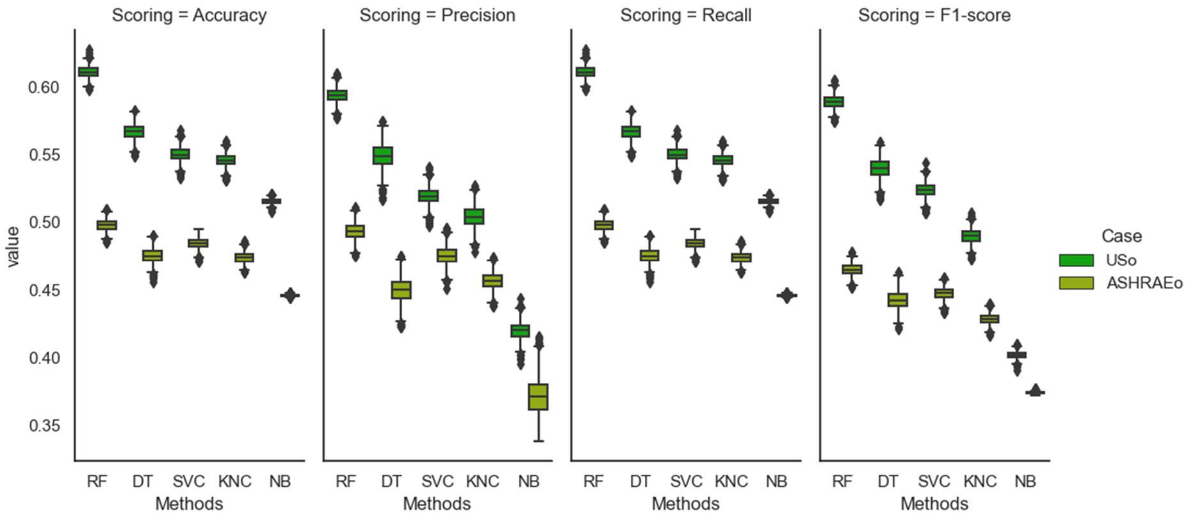

Figure 10. Comparison of classification models based on scoring, with own variables in the US database (USo) and ASHRAE database (ASHRAEo).

KNC, SVC, and NB models have different characteristics and assumptions compared to RF and DT models. In KNC, the classification decision is based on the similarity of a data point to its K-nearest neighbours. In this case, adding more variables increases the dimensionality of the feature space, which can lead to increased sparsity and make it harder to find meaningful neighbours. This can result in a decrease in accuracy, precision, recall, and F1-score as the number of variables increases. In the case of the US, from 11 to 21 variables, and in the case of the ASHRAE database, increase in 1 variable, for this reason, the variation in this case is smaller or none.

In the case of SVC, on the other hand, it aims to find a hyperplane that separates different classes in a high-dimensional feature space. As the number of variables increases, the complexity of finding an optimal hyperplane also increases. This can lead to reduced performance, especially if the number of variables exceeds the available training instances or when the added variables do not contribute meaningful discriminatory information. Consequently, accuracy, precision, recall, and F1-score may decrease with an increased number of variables in SVC. On the other hand, NB models assume feature independence, meaning that they suppose that the variables are conditionally independent given the class. If the additional variables introduce dependencies or break this assumption, it can negatively impact the performance of the NB model.

In this study, the importance of variables in the estimation of indoor thermal comfort refers to understanding which factors have a significant impact on individuals' perception of comfort in indoor environments. The identification and consideration of these variables are essential for the design and optimisation of HVAC system strategies and occupant-orientated thermal comfort models.

It is also important to mention that the impact of the number of variables on model performance is highly dependent on the dataset, and as shown in this study, the main predictor variables represent the specific characteristics of the TSV problem.

In practise, it is advisable to conduct experiments to analyse cultural and climate aspects in different regions and to assess the effect of variable selection on the performance of different models in their own context.

To accurately estimate thermal comfort, it is often necessary to consider several variables simultaneously and apply adaptive comfort models that take into account the dynamic interaction between the occupants and the environment. By understanding the importance of these variables, building designers, facility managers, and HVAC engineers can make informed decisions to create more comfortable indoor environments, while optimising energy efficiency and occupant satisfaction.

5.3. Thermal Adaptation and the Importance of Variables

This subsection highlights the concept of thermal adaptation and its three types: physiological, psychological, and behavioural adaptation and the relationship with variables.

These theoretical aspects provide a framework for understanding the complex nature of thermal comfort and how individuals adapt to different indoor thermal conditions. By incorporating these concepts into the interpretation of this analysis, it was possible to contextualise the findings and gain a deeper understanding of the factors influencing indoor thermal sensation.

Physiological adaptation, which refers to the body's physical response to changes in temperature, emerges as a key variable in the analysis. Specifically, indoor, and operating temperatures were found to have a significant impact on predicting the thermal sensation. These variables reflect the immediate thermal conditions of the indoor environment, indicating the importance of maintaining appropriate temperatures for occupant comfort. In addition, secondary variables related were identified, highlighting the influence of outdoor conditions on indoor thermal perception, and with less influence, the humidity and the max/min outdoor temperature.

Psychological adaptation, influenced by previous experiences, is captured in the analysis through the inclusion of the running mean temperature and humidity. These variables provide an indication of the occupants' long-term exposure to certain thermal conditions, affecting their perception and adaptation to indoor temperatures.

Behavioural adaptation, as outlined by Brager and de Dear [60], includes the adjustments individuals make to their behaviour to achieve thermal comfort. In this study, the main variable representing behavioural adaptation was clothing insulation. This variable captures the choices individuals make in terms of clothing layers, which can significantly affect their perception of thermal comfort. This highlights the importance of research to develop effective strategies to improve occupant satisfaction with clothing advice systems.

After analysing the variables associated with different types of thermal adaptation, we gain insights into the factors that significantly influence indoor thermal sensation, which is necessary for future studies.

6. Conclusions

There is a wide range of variables that can be used to predict the thermal sensation of building occupants through algorithms to optimise their thermal comfort with lower energy consumption, developing actions that improve the environment for users. These predictions, made with learning techniques, can be used to make decisions in the conditioning systems in terms of sustainability and economy, with them having the variables that influence the behaviour of the occupants to adapt the environment to their thermal preferences as a main contribution to generating a positive impact on energy savings in buildings. On the other hand, it is necessary to know the influence of different features in comfort studies for future considerations in the design phases of comfort studies in buildings.

Through a comprehensive study of indoor thermal comfort in office buildings in Mediterranean climates, this paper presents a robust methodology using five different techniques and two different databases. The research results culminate in a number of key findings that not only underline the importance of specific variables but also highlight the effectiveness of machine learning in this field.

The research has led to the following conclusions:

- Fundamental variables: The analysis identifies eight key variables that determine indoor thermal comfort in Mediterranean climates. These variables include indoor air temperature, mean radiant temperature, indoor globe temperature, CO_2, age, outdoor temperature, indoor humidity, and running mean temperature.
- Extended dataset: Expanding the variables considered from eight key variables to a broader set of 21 variables improves the accuracy and quality of the metrics. This expansion demonstrates a 5% increase in the effectiveness of the techniques with respect to the fundamental variables.

- Machine learning empowerment: Research demonstrates the potential of these strategies and emphasises the importance of machine learning algorithms in understanding the thermal comfort of the users. Among the techniques examined, random forest stands out as the one that achieves the greatest metrics, demonstrating improved accuracy and performance overall.
- Best performer: Throughout the research, random forest consistently outperforms alternative machine learning techniques. This is shown for both the experimental dataset from the University of Seville and the Mediterranean climate dataset from the ASHRAE database.
- Future research and implementation: Further studies based on the Mediterranean climate should include larger datasets to increase the robustness of the results. Furthermore, translating the present results into practical applications has the potential to revolutionise building design, operational strategies, and adaptive measures to optimise occupant thermal comfort.

This research has demonstrated an advance in the understanding of indoor thermal comfort in Mediterranean office buildings. The effectiveness of the methodology, the importance of the key variables and the analysis of machine learning techniques, in particular random forest, provide a basis for future research and transformative real-world applications.

Author Contributions: Conceptualisation, P.A.-R. and J.G.; methodology, P.A.-R. and E.B.-M.; software, P.A.-R. and J.N.; validation, E.B.-M. and J.G.; formal analysis, E.B.-M.; investigation, P.A.-R. and E.B.-M.; data curation, E.B.-M.; writing—original draft preparation, P.A.-R.; writing—review and editing, P.A.-R. and E.B.-M.; visualisation, P.A.-R.; supervision, J.G.; project administration, P.A.-R. and J.G.; funding acquisition, P.A.-R. and J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Consejería de Economía, Conocimiento, Empresas y Universidad, Junta de Andalucía, grant number US-1380581 through the project SICODE—"Sistema Inteligente de Confort para Optimizar la Demanda Energética en edificios terciarios".

Data Availability Statement: The ASHRAE dataset used can be freely downloaded from the ASHRAE Global Thermal Comfort Database II (https://datadryad.org/stash/dataset/doi:10.6078/D1F671 (accessed on 19 July 2023), file 'ashrae_db2.01.csv').

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ASHRAE	American Society of Heating, Refrigerating and Air-conditioning engineers
BP	Blood pressure
BT	Begged trees
CC	Climate chambers
CLO	Clothing insulation
DT	Decision tree
FN	False negatives
FP	False positives
GPC	Gaussian process classifier
GS	Grid search
HVAC	Heating, ventilating, and air conditioning
ISD	Integrated surface dataset
KNC	K nearest neighbour classifier
MET	Metabolic rate
ML	Machine learning
MM	Mixed mode
NB	Naive Bayes
NN	Neural network
NL	Naturally ventilated

NS	Non-specific
O	Office building
PR	Pulse rate
LDA	Linear discriminant analysis
LR	Logistic regression
RF	Random forest
RH	Relative humidity
RS	Random search
ST	Skin temperature
SVC	Support vector machine classifier
TN	True negatives
TP	True positives
TSV	Thermal sensation vote
U	University buildings

References

1. Fanger, P.O. *Analysis and Applications in Environmental Engineering*; Danish Technical Press: Copenhagen, Denmark, 1970; ISBN 0070199159.
2. Nicol, J.F.; Humphreys, M.A. Adaptive Thermal Comfort and Sustainable Thermal Standards for Buildings. *Energy Build.* **2002**, *34*, 563–572. [CrossRef]
3. Zhao, Q.; Lian, Z.; Lai, D. Thermal Comfort Models and Their Developments: A Review. *Energy Built Environ.* **2021**, *2*, 21–33. [CrossRef]
4. *55-2020*; Thermal Environmental Conditions for Human Occupancy. American Society of Heating, Refrigerating and Air conditioning Engineers (ASHRAE): Atlanta, GA, USA, 2020.
5. Fard, Z.Q.; Zomorodian, Z.S.; Korsavi, S.S. Application of Machine Learning in Thermal Comfort Studies: A Review of Methods, Performance and Challenges. *Energy Build.* **2022**, *256*, 111771. [CrossRef]
6. Luo, M.; Xie, J.; Yan, Y.; Ke, Z.; Yu, P.; Wang, Z.; Zhang, J. Comparing Machine Learning Algorithms in Predicting Thermal Sensation Using ASHRAE Comfort Database II. *Energy Build.* **2020**, *210*, 109776. [CrossRef]
7. Yang, B.; Li, X.; Liu, Y.; Chen, L.; Guo, R.; Wang, F.; Yan, K. Comparison of Models for Predicting Winter Individual Thermal Comfort Based on Machine Learning Algorithms. *Build. Environ.* **2022**, *215*, 108970. [CrossRef]
8. Ma, N.; Aviv, D.; Guo, H.; Braham, W.W. Measuring the Right Factors: A Review of Variables and Models for Thermal Comfort and Indoor Air Quality. *Renew. Sustain. Energy Rev.* **2021**, *135*, 110436. [CrossRef]
9. Mamani, T.; Herrera, R.F.; Muñoz-La Rivera, F.; Atencio, E. Variables That Affect Thermal Comfort and Its Measuring Instruments: A Systematic Review. *Sustainability* **2022**, *14*, 1773. [CrossRef]
10. Xie, J.; Li, H.; Li, C.; Zhang, J.; Luo, M. Review on Occupant-Centric Thermal Comfort Sensing, Predicting, and Controlling. *Energy Build.* **2020**, *226*, 110392. [CrossRef]
11. Čulić, A.; Nižetić, S.; Šolić, P.; Perković, T.; Čongradac, V. Smart Monitoring Technologies for Personal Thermal Comfort: A Review. *J. Clean. Prod.* **2021**, *312*, 127685. [CrossRef]
12. Liu, S.; Schiavon, S.; Das, H.P.; Jin, M.; Spanos, C.J. Personal Thermal Comfort Models with Wearable Sensors. *Build. Environ.* **2019**, *162*, 106281. [CrossRef]
13. Lu, S.; Wang, W.; Wang, S.; Cochran Hameen, E. Thermal Comfort-Based Personalized Models with Non-Intrusive Sensing Technique in Office Buildings. *Appl. Sci.* **2019**, *9*, 1768. [CrossRef]
14. Wang, Z.; Liu, J.; Zhang, Y.; Yuan, H.; Zhang, R.; Srinivasan, R.S. Practical Issues in Implementing Machine-Learning Models for Building Energy Efficiency: Moving beyond Obstacles. *Renew. Sustain. Energy Rev.* **2021**, *143*, 110929. [CrossRef]
15. Jin, L.; Liu, T.; Ma, J. Modeling Thermal Sensation Prediction Using Random Forest Classifier. *Commun. Comput. Inf. Sci.* **2021**, *1469*, 552–561.
16. Li, Q.Y.; Han, J.; Lu, L. A Random Forest Classification Algorithm Based Personal Thermal Sensation Model for Personalized Conditioning System in Office Buildings. *Comput. J.* **2021**, *64*, 500–508. [CrossRef]
17. Lu, S.; Wang, W.; Lin, C.; Hameen, E.C. Data-Driven Simulation of a Thermal Comfort-Based Temperature Set-Point Control with ASHRAE RP884. *Build. Environ.* **2019**, *156*, 137–146. [CrossRef]
18. Salleh, F.H.M.; binti Saripuddin, M.; bin Omar, R. Predicting Thermal Comfort of HVAC Building Using 6 Thermal Factors. In Proceedings of the 2020 8th International Conference on Information Technology and Multimedia (ICIMU), Selangor, Malaysia, 24–25 August 2020; pp. 170–176.
19. Parkinson, T.; de Dear, R.; Brager, G. Nudging the Adaptive Thermal Comfort Model. *Energy Build.* **2020**, *206*, 109559. [CrossRef]
20. Aryal, A.; Becerik-Gerber, B. Thermal Comfort Modeling When Personalized Comfort Systems Are in Use: Comparison of Sensing and Learning Methods. *Build. Environ.* **2020**, *185*, 107316. [CrossRef]
21. Chaudhuri, T.; Zhai, D.; Soh, Y.C.; Li, H.; Xie, L. Random Forest Based Thermal Comfort Prediction from Gender-Specific Physiological Parameters Using Wearable Sensing Technology. *Energy Build.* **2018**, *166*, 391–406. [CrossRef]

22. Cosma, A.C.; Simha, R. Machine Learning Method for Real-Time Non-Invasive Prediction of Individual Thermal Preference in Transient Conditions. *Build. Environ.* **2019**, *148*, 372–383. [CrossRef]
23. Jiang, L.; Yao, R. Modelling Personal Thermal Sensations Using C-Support Vector Classification (C-SVC) Algorithm. *Build. Environ.* **2016**, *99*, 98–106. [CrossRef]
24. Song, G.; Ai, Z.; Zhang, G.; Peng, Y.; Wang, W.; Yan, Y. Using Machine Learning Algorithms to Multidimensional Analysis of Subjective Thermal Comfort in a Library. *Build. Environ.* **2022**, *212*, 108790. [CrossRef]
25. Xiong, L.; Yao, Y. Study on an Adaptive Thermal Comfort Model with K-Nearest-Neighbors (KNN) Algorithm. *Build. Environ.* **2021**, *202*, 108026. [CrossRef]
26. Chaudhuri, T.; Soh, Y.C.; Li, H.; Xie, L. Machine Learning Based Prediction of Thermal Comfort in Buildings of Equatorial Singapore. In Proceedings of the 2017 IEEE International Conference on Smart Grid and Smart Cities, ICSGSC 2017, Singapore, 23–26 July 2017; pp. 72–77.
27. Iyoho, A.E.; Ng, L.J.; MacFadden, L. Modeling of Gender Differences in Thermoregulation. *Mil. Med.* **2017**, *182*, 295–303. [CrossRef] [PubMed]
28. Schweiker, M.; Fuchs, X.; Becker, S.; Shukuya, M.; Dovjak, M.; Hawighorst, M.; Kolarik, J. Challenging the Assumptions for Thermal Sensation Scales Challenging the Assumptions for Thermal Sensation Scales. *Build. Res. Inf.* **2016**, *45*, 572–589. [CrossRef]
29. Subasi, A. Machine Learning Techniques. In *Practical Machine Learning for Data Analysis Using Python*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 91–202.
30. Cutler, A.; Cutler, D.R.; Stevens, J.R. Random Forests. In *Ensemble Machine Learning*; Springer: Boston, MA, USA, 2012; pp. 157–175.
31. Altan, H.; Ozarisoy, B. Dynamic Evaluation Method for Assessing Households' Thermal Sensation Using Parametric Statistical Analysis: S Longitudinal Field Study in the South-Eastern Mediterranean Climate. In Proceedings of the CLIMA 2022 Conference, Rotterdam, The Netherlands, 22–25 May 2022. [CrossRef]
32. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In *Encyclopedia of Database Systems*; Springer: Boston, MA, USA, 2009; pp. 532–538.
33. Ozarisoy, B.; Altan, H. Regression Forecasting of 'Neutral' Adaptive Thermal Comfort: A Field Study Investigation in the South-Eastern Mediterranean Climate of Cyprus. *Build. Environ.* **2021**, *202*, 108013. [CrossRef]
34. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. [CrossRef]
35. Kazemi, F.; Asgarkhani, N.; Jankowski, R. Machine Learning-Based Seismic Response and Performance Assessment of Reinforced Concrete Buildings. *Arch. Civ. Mech. Eng.* **2023**, *23*, 94. [CrossRef]
36. Bagherzadeh, F.; Shafighfard, T.; Khan, R.M.A.; Szczuko, P.; Mieloszyk, M. Prediction of Maximum Tensile Stress in Plain-Weave Composite Laminates with Interacting Holes via Stacked Machine Learning Algorithms: A Comparative Study. *Mech. Syst. Signal Process.* **2023**, *195*, 110315. [CrossRef]
37. Yang, L.; Shami, A. On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice. *Neurocomputing* **2020**, *415*, 295–316. [CrossRef]
38. Belete, D.M.; Huchaiah, M.D. Grid Search in Hyperparameter Optimization of Machine Learning Models for Prediction of HIV/AIDS Test Results. *Int. J. Comput. Appl.* **2022**, *44*, 875–886. [CrossRef]
39. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
40. Wu, J.; Chen, X.Y.; Zhang, H.; Xiong, L.D.; Lei, H.; Deng, S.H. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *J. Electron. Sci. Technol.* **2019**, *17*, 26–40. [CrossRef]
41. Barbadilla-Martín, E.; Salmerón Lissén, J.M.; Guadix Martín, J.; Aparicio-Ruiz, P.; Brotas, L. Field Study on Adaptive Thermal Comfort in Mixed Mode Office Buildings in Southwestern Area of Spain. *Build. Environ.* **2017**, *123*, 163–175. [CrossRef]
42. Beck, H.E.; Zimmermann, N.E.; McVicar, T.R.; Vergopolan, N.; Berg, A.; Wood, E.F. Present and Future Köppen-Geiger Climate Classification Maps at 1-Km Resolution. *Sci. Data* **2018**, *5*, 180214. [CrossRef] [PubMed]
43. Arakawa Martins, L.; Soebarto, V.; Williamson, T. A Systematic Review of Personal Thermal Comfort Models. *Build. Environ.* **2022**, *207*, 108502. [CrossRef]
44. Nicol, F.; Humphreys, M. Derivation of the Adaptive Equations for Thermal Comfort in Free-Running Buildings in European Standard EN15251. *Build. Environ.* **2010**, *45*, 11–17. [CrossRef]
45. Humphreys, M.A.; Rijal, H.B.; Nicol, J.F. Updating the Adaptive Relation between Climate and Comfort Indoors; New Insights and an Extended Database. *Build. Environ.* **2013**, *63*, 40–55. [CrossRef]
46. de la Hoz-Torres, M.L.; Aguilar, A.J.; Costa, N.; Arezes, P.; Ruiz, D.P.; Martínez-Aires, M.D. Predictive Model of Clothing Insulation in Naturally Ventilated Educational Buildings. *Buildings* **2023**, *13*, 1002. [CrossRef]
47. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
48. Kotsiantis, S.B. Decision Trees: A Recent Overview. *Artif. Intell. Rev.* **2013**, *39*, 261–283. [CrossRef]
49. Salcedo-Sanz, S.; Rojo-Álvarez, J.L.; Martínez-Ramón, M.; Camps-Valls, G. Support Vector Machines in Engineering: An Overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2014**, *4*, 234–267. [CrossRef]
50. Zhang, H. Exploring Conditions for the Optimality of Naïve Bayes. *Int. J. Pattern Recognit. Artif. Intell.* **2005**, *19*, 183–198. [CrossRef]
51. Luo, M. *The Dynamics and Mechanism of Human Thermal Adaptation in Building Environment—A Glimpse to Adaptive Thermal Comfort in Buildings*; Springer: Berlin/Heidelberg, Germany, 2020; ISBN 978-981-15-1164-6.

52. Seyedzadeh, S.; Rahimian, F.P.; Rastogi, P.; Glesk, I. Tuning Machine Learning Models for Prediction of Building Energy Loads. *Sustain. Cities Soc.* **2019**, *47*, 101484. [CrossRef]
53. Kottek, M.; Grieser, J.; Beck, C.; Rudolf, B.; Rubel, F. World Map of the Köppen-Geiger Climate Classification Updated. *Meteorol. Z.* **2006**, *15*, 259–263. [CrossRef] [PubMed]
54. Baker, N.; Standeven, M. Comfort Criteria for Passively Cooled Buildings a Pascool Task. *Renew. Energy* **1994**, *5*, 977–984. [CrossRef]
55. Benton, C.O. *Sunset Building: A Study of Occupant Thermal Comfort in Support of PG&E'S Advanced Customer Technology Test(ACT2) or Maximum Energy Efficiency*; Final Report, CEDR-06-94, CEDR; University of California Berkeley: Berkeley, CA, USA, 1994; p. 94.
56. Brager, G.; Paliaga, G.; De Dear, R. Operable Windows, Personal Control, and Occupant Comfort. *ASHRAE Trans.* **2004**, *110*, 17–35.
57. Stoops, J.L. The Thermal Environmentand Occupant Perceptions in European Office Buildings. Ph.D. Thesis, Chalmers University of Technology, Gothenburg, Sweden, 2000.
58. Zhou, X.; Xu, L.; Zhang, J.; Niu, B.; Luo, M.; Zhou, G.; Zhang, X. Data-Driven Thermal Comfort Model via Support Vector Machine Algorithms: Insights from ASHRAE RP-884 Database. *Energy Build.* **2020**, *211*, 109795. [CrossRef]
59. Liu, X.; Tang, H.; Ding, Y.; Yan, D. Investigating the Performance of Machine Learning Models Combined with Different Feature Selection Methods to Estimate the Energy Consumption of Buildings. *Energy Build.* **2022**, *273*, 112408. [CrossRef]
60. De Dear, R.; Brager, G. Developing an Adaptive Model of Thermal Comfort and Preference. *ASHRAE Trans.* **1998**, *104*, 145–167.