Solar Energy 218 (2021) 95–107

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Feature selection for CIE standard sky classification

Diego Granados-López [a], Andrés Suárez-García [a,b], Montserrat Díez-Mediavilla [a], Cristina Alonso-Tristán [a,*]

[a] *Research Group Solar and Wind Feasibility Technologies (SWIFT), Electromechanical Engineering Department, Universidad de Burgos, 09006 Burgos, Spain*
[b] *Centro Universitario de la Defensa. Escuela Naval Militar de Marín, 36920 Marin, Pontevedra, Spain*

ARTICLE INFO

Keywords:
CIE standard sky classification
Feature selection
Meteorological indices
Machine learning

ABSTRACT

There are several compilations of sky classifications that refer to Meteorological Indices (MIs) (variables usually recorded at meteorological ground stations), due to the scarcity of sky scanner devices that can supply the experimental data needed to apply the CIE standard sky classification. The use of one rather than another MI is never justified, because there is no standardized criterion for their selection. In this study, forty-three MIs, traditionally used to define different sky conditions, are reviewed. Feature Selection (FS) is a key step in the design of a sky-classification algorithm using MIs as an alternative to data from sky scanners. Four procedural methods for FS -Pearson, Permutation Importance, Recursive Feature Elimination, and Boruta- are applied to an extensive data set of MIs that includes CIE standard sky classification data, which was used as a reference. The use of FS procedures significatively reduced the original set of MIs, permitting the construction of different classification trees with high performance for the sky classification. In the case of the Pearson FS method, the classification tree only used two MIs. The advantage of the Pearson FS method is that it functions independently from the machine-learning algorithm used latter for the sky classification.

1. Introduction

Daylight, as part of the solar irradiance is an essential natural resource even for human health. In building design, projection of daylight can increase the energy efficiency of buildings (Dubois et al., 2016; Fouquart et al., 1990; Li, 2010) and will have positive effects on the well-being of occupants (Aries et al., 2015; Edwards and Torcellini, 2002). Natural lighting availability is highly dependent on luminance levels and sky conditions. In 2003, the Commission International de L'Eclairage (CIE) (Uetani et al., 2003) and the International Organization for Standardization (ISO) (ISO, 2004) both adopted 15 Standard Sky Luminance Distributions (SSLD), five clear, five overcast and five partly cloudy sky conditions. They provide the most versatile definition of skylight at various localities and daylight climate regions, making it possible to simulate an annual daylight profile at any point on earth in absolute units, based on typical luminance sky patterns.

The sky scanner is the standard instrument for measuring sky luminance distribution (Li, 2010). Despite the high interest in those measurements, very few studies at only a handful of European (Markou et al., 2005; Markou et al., 2004; Suárez-García et al., 2018; Torres et al., 2010a, b; Tregenza, 1999) and Asian (Chaiwiwatworakul and Chirarattananon, 2004; Li and Tang, 2008; Ng et al., 2007; Zi et al., 2020) locations have been conducted to characterize the sky under the CIE standard, mainly due to the scarcity of sky scanner devices available to gather sky luminance data.

Different alternatives to the use of sky scanners have been proposed for classifying the skies (Li et al., 2014b), including the use of different climatic parameters or meteorological indices (Li et al., 2004; Lou et al., 2017; Umemiya and Kanou, 2008; Wong et al., 2012), vertical (Chen et al., 2019; Darula et al., 2013) and horizontal illuminance data (Alshaibani, 2016b; Alshaibani, 2017), as well as satellite data (Janjai et al., 2008). Added difficulties for sky classification (Allard et al., 2015) include the variability of sky conditions and their sensitivity to many stochastic variables.

Sky classification in various studies relies on Meteorological Indices (MIs), factors usually recorded at meteorological ground stations that, to a greater or lesser extent, affect the luminance and radiance distributions: sun position, cloud coverage, turbidity, and weather conditions, among others. Such climatic indices within certain ranges will lead to sky luminance and radiance distributions with similar features, and a straightforward approach is to describe those distributions by sky conditions (Lou et al., 2017). The selection of the MIs depends on their

[*] Corresponding author.
E-mail addresses: catristan@ubu.es, cristinaalonso.tristan@gmail.com (C. Alonso-Tristán).

https://doi.org/10.1016/j.solener.2021.02.039
Received 15 September 2020; Received in revised form 24 November 2020; Accepted 15 February 2021
Available online 8 March 2021

availability. The number of MIs used and the conditions that define each sky type vary in each study, complicating the task of defining a taxonomy that could unequivocally describe the specific characteristics of each sky type (Dieste-Velasco et al., 2019; Perez et al., 1990a), even in a reduced classification with only three sky types: clear, partly cloudy, and overcast conditions.

In recent years, machine learning (ML) tools, such as Artificial Neural Networks (ANN's) (Li et al., 2010) and classification trees have, among others, been applied to sky classification. Supervised machine learning tools permit the identification of patterns and relationships between inputs and outputs, as long as the algorithm has sufficient examples to train recognition. In this paper, a set of sky type samples labeled as CIE Standard Sky Classification will be used as the training set for sky classification purposes and a test set of available MIs will be used as input for the algorithm.

The work flow of a supervised Machine-Learning (ML) tool is shown in Fig. 1. The first step for every ML tool is to filter and to analyze the input data so as categorize it and to control its quality. The second step is the Feature Selection (FS) procedure: the identification of related features within a set of data and the removal of irrelevant or less important features that contribute little or nothing to the definition of the target variable, so as to achieve models of greater accuracy. FS is one of the core concepts of ML that will impact on the performance of the developed model, improving its precision and reducing its complexity and overfitting as well as its runtime.

Following acceptable FS, the algorithm is trained using part of the input data set (training set), typically 80% of the total, using the remaining 20% for validation tests. Training set data and test set data are exchanged as many times as necessary, until the algorithm is considered validated.

In this study, a total of forty-three MIs describing sky conditions are borrowed from past studies for their use as variables to define sky types. The use of one rather than another MI is not justified, because there is no standardized criterion for selecting MIs. It is necessary to compare the information that each of them offers, removing those that offer redundant or insignificant information for the definition of sky types (Yang and Pedersen, 1997). Therefore, FS is a key step in the design of a sky classification algorithm using MIs as an alternative to data from sky scanners. The main objective of this study is to determine, through a FS procedure, the most suitable MIs and their precise number for the optimization of the sky classification algorithms. Forty-three MIs were included in the study, calculated from half-hourly experimental data records collected at Burgos, Spain, between September 2016 and December 2019. The following FS criteria were selected: Pearson (Biesiada and Duch, 2007), Permutation Importance (Gregorutti et al., 2017), Recursive Feature Elimination, and Boruta (Degenhardt et al., 2019).

This study reports an extensive review of the MIs that define different sky conditions and features that are suitable for sky classification. Structured and rigorous FS procedures can determine the usefulness of the information in these indices, with a high degree of success, for the problem of sky classification, the informative equivalence between some of the MIs, and the number of MIs that may be needed for sky classification in line with the CIE standard. It was proven that the Pearson FS procedure performed accurate sky classification into three sky conditions (clear, partly cloudy and overcast conditions), in accordance with the CIE Standard Classification, requiring only two MIs. The FS results, processed in a classification tree to test their validity, confirmed that the intervals of definition of the MIs for each sky type were close to the intervals that were established in each study for the individual use of each MI.

The structure of this paper will be as follows. Following the Introduction in Section 1, the methodology will be explained in Section 2, where the experimental facility and the data processing needed to calculate the MIs and the experimental campaign is introduced in Section 2.1. In Section 2.2, the CIE Standard Sky classification of Burgos, Spain, gathered during the experimental campaign will be described, as reference data for sky classification. The MIs with their data on sky conditions that were available for the FS procedures will be reviewed in Section 2.3. Then, the different FS procedures used in this work and the results of their application to the experimental MIs will be described in Section 2.4. In Section 2.5, the classification trees will be introduced, together with the machine learning algorithm used to test the performance of the FS procedure; and in Section 2.6 the metrics used to test FS performance will be presented. Finally, the main results and the conclusions of the study will be summarized in Sections 3 and 4.

2. Methodology

The present work was developed in four steps: data collection, Feature Selection (FS), classification trees, and classification metrics. Several meteorological variables were collected between 21 September 2016 and 31 January 2020. The dataset contained over eight-thousand samples that were used for the evaluation of 43 MIs. The size of the data set lent support for the conclusions of this work. Following the calculation of the MIs, the classification tree was employed in conjunction with the FS procedure to classify sky cloudiness (clear, partial or overcast) following the established CIE patterns. The ML classification tree algorithm was selected, because it can process and extract the rules for sky labelling. FS, for maximum simplification of the classification trees, was applied, in an effort to reduce the number of MIs serving as ML algorithm inputs to a minimum. Finally, the outputs of the classification tree algorithm were analyzed using several metrics.

2.1. The experimental facility

The experimental campaign during which the meteorological data were recorded for the processing of each MI in this study was performed

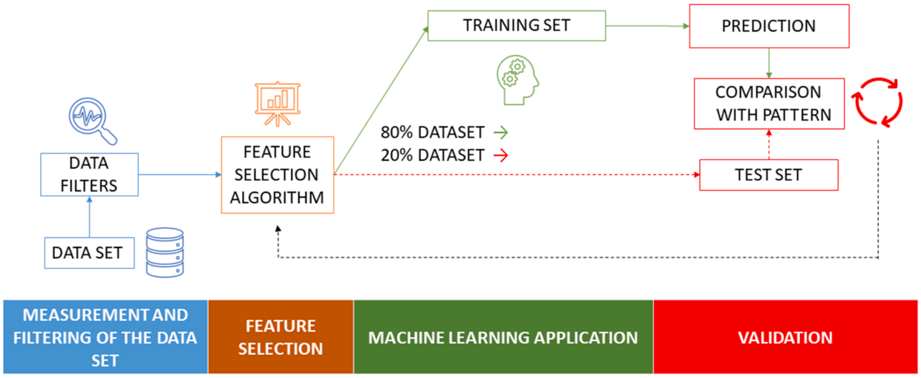

Fig. 1. Workflow of Supervised Machine-Learning tool.

in Burgos, Spain. Data collection took place at a meteorological facility located at the Higher Polytechnic School of Burgos University (LON. 42°21′04″N, LAT. 3°41′20″W, 856 m above mean sea level). A commercial sky scanner from Eko instruments, model MS-321LR, was used for CIE Standard classification. Its technical specifications are shown in Table 1. Measurements from 4-minute scans were taken every 15 min, from September 2016 to December 2017. From January 2018, the scans were taken every 10 min. The device was adjusted on a monthly basis to measure from sunrise to sunset. First and last daily records were discarded, to avoid measurements with solar altitudes equal to or lower than 7.5°. Data higher than 50 kcd/m^2 and lower than 0.1 kcd/m^2 were also discarded, following the technical specifications of the sky scanner. Seven lux sensors, EKO, model ML-020S-O, technical specification listed in Table 2, were also used: four of them recorded vertical global illuminance in the four cardinal orientations and three lux sensors recorded horizontal, global, beam, and diffuse illuminance. Horizontal global, diffuse and beam irradiance were measured using Hukseflux pyranometers, model SR11 and a Hukseflux pyrheliometer, model DR01. The technical specifications of the pyranometers and the pyrheliometer are shown in Table 3. The beam illuminance and irradiance sensors were installed on a sun tracker, model Sun-Tracker 3000, from Geónica. The diffuse illuminance and irradiance sensors were obscured from direct sunlight by a shadow hat. Illuminance and irradiance data were recorded every 10 min (averaging recorded scans of 30 s). Fig. 2 shows the experimental equipment.

CIE quality criteria (Comission Internationale de, L.E, 1995) were used for analyze and filter illuminance data while irradiance data were analyzed and then filtered using conventional quality criteria (Gueymard and Ruiz-Arias, 2016). To match simultaneous records of illuminance and irradiance data, half-hourly and hourly sky scanner measurements were used in this study, from September 2016 to December 2017, and from January 2018, ten minutes records. If the illuminance and irradiance data failed to pass the quality criteria, then all the simultaneous data sets were rejected.

The measurement campaign extended between 21 September 2016, and 31, January 2020. Following their analysis and the filtering process, the experimental data amounted to 8829 items.

2.2. CIE standard classification of Burgos skies

Supervised Machine Learning needs examples for training the classifier algorithm. In this work, the CIE standard classification served as a benchmark for estimating the performance of the supervised machine learning algorithm and for testing the FS procedure. Several works have reported that the CIE Standard sky classification provides a good overall framework for representing the actual conditions for homogeneous skies (Li et al., 2011b; Li et al., 2004; Li et al., 2010; Markou et al., 2005). Tregenza (Tregenza, 2004) gave a detailed description of the CIE standard classification procedure following a discrete integration methodology, the same method that was used for sky classification in Burgos. The labelling of CIE sky types was as follows: I.1 to III.1: cloudy; III.2 to IV.3: partially cloudy; and, IV.4 to VI.6: clear skies. More information on the classification method can be obtained from a previous work (Suárez-García et al., 2018). In the experimental campaign between 21 September 2016, and 31 January 2020, clear skies predominated in Burgos (52%) while overcast skies were present in 15% and partially cloudy skies in 33% of cases, as shown in Fig. 3.

Table 1
Sky scanner specifications.

Model	MS-321LR Sky Scanner
Dimensions (W × D × H)	430 mm × 380 mm × 440 mm
Mass	12.5 kg
FOV	11°
Luminance	0 to 50 kcd/m^2
Radiance	0 to 300 W/m^2
A/D Convertor	16 bits
Calibration Error	2%

Table 2
Luxmeter technical specifications.

Model	ML-020S-O
Illuminance Range	0 to 150,000 lx
Output	0 to 30,000 μV
Impedance	280 Ω
Operating temperature range	−10 °C to 50 °C
Temperature response	0.4%

Table 3
Pyranometers and Pyrheliometer technical specifications.

Model	SR11	DR01
Measurement Range	0–3000 W/m^2	0–4000 W/m^2
Calibration uncertainty	<1.8% (k = 2)	<1.2% (k = 2)
Spectral Range	285–3000 × 10^{-9} m	200–4000 × 10^{-9}m
Sensitivity (nominal)	15 × 10^{-6} V/(W/m^2)	10 × 10–6 V/(W/m^2)
Operating temperature range	−40 °C to 80 °C	−40 °C to 80 °C
Temperature response	<± 2% (−10 °C to 40 °C)	<± 1% (−10 to +40 °C)

2.3. Meteorological indices

Skies of the same category are assumed to share identical well-defined sky luminance patterns (Darula and Kittler, 2002), which is the straightforward approach for sky classification. Once the skies have been identified, the daylight on any surface can be estimated, by integrating the luminance distribution of the sky dome over each surface (Granados-López et al., 2020). Therefore, any climatic parameter based on lighting measurements can potentially identify a given sky condition. Table 4 describes the 43 MIs reviewed in this work.

The US National Bureau of Standards (NBS) recommends the use of the horizontal diffuse fraction, k_d, for sky classification (Fakra et al., 2011): low k_d values indicate clear sky conditions and high values are usually present, but not exclusively so, in overcast conditions (Li et al., 2015). Alternatively, high values of the horizontal direct fraction, k_b, are representative of clear skies, due to the high values of the solar irradiation beam component (Ferraro et al., 2010) while low k_b values predominate on cloudy days,.

Perez's Sky clearness, ε_p, maybe one of the most widely used MIs for sky characterization, was originally proposed to define the ratio of illuminance and irradiance, known as luminous efficacy, K. The sky's brightness index, Δ, is often used with the clearness index, ε_p, for sky classification (Li et al., 2004; Perez et al., 1990a).

Luminous efficacy, K, can be modeled through different parameters such as the solar zenith angle, Z_s; Perez's sky clearness index, ε_p; the sky's brightness index, Δ, and, the atmospheric precipitable water content (Perez et al., 1990a).

Relative heaviness, Ω, (Chung, 1992) is proportional to the amount of solar radiation entering into clouds. Cloud cover, CC, is often used as an indicator of sky conditions (Muneer et al., 2007): 0 oktas is the CC value for clear skies and 8 oktas is assigned in overcast conditions. A sky classification based on MIs was performed by Igawa *et al.* (Igawa et al., 2004) using Igawa's sky index, S_i, the clear sky index, k_c, and the cloudless index, Cle.

A CIE-based standard classification of skies using global horizontal illuminance, $LxGH$, and Kittler's index, k_t, was proposed by Lou et al. (2019). Kittler's index, k_t, is widely used for illumination studies, due to the high information content that it provides when only global irradiation data are available. However, k_t is only available when the zenith sun angle is under 80°, $Z_s < 80^{\circ}$. An alternative and globally valid definition was proposed by Perez et al. (1990b), k_{t2}, that used cloud cover, CC, together with relative humidity, RH, among other factors,

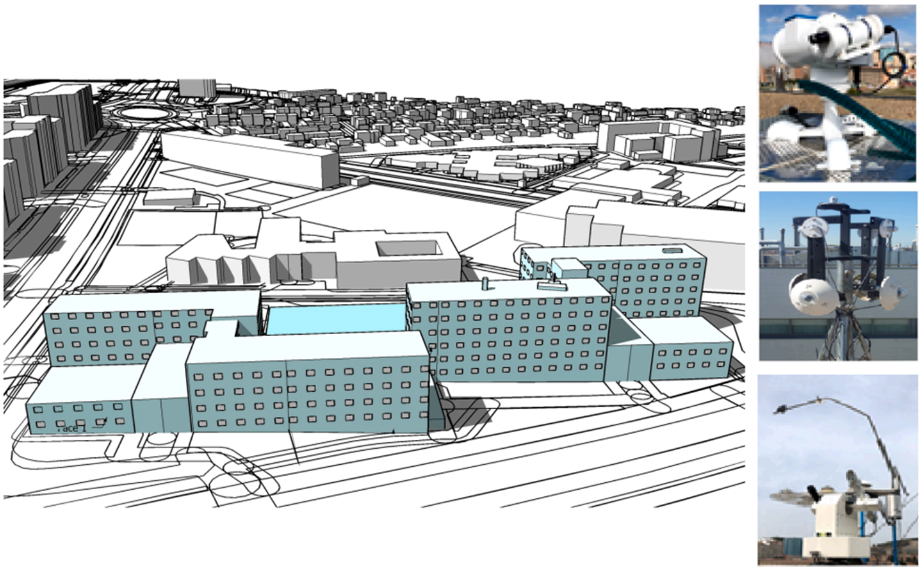

Fig. 2. Experimental equipment on the roof of the Higher Polytechnic School of Burgos University, Spain.

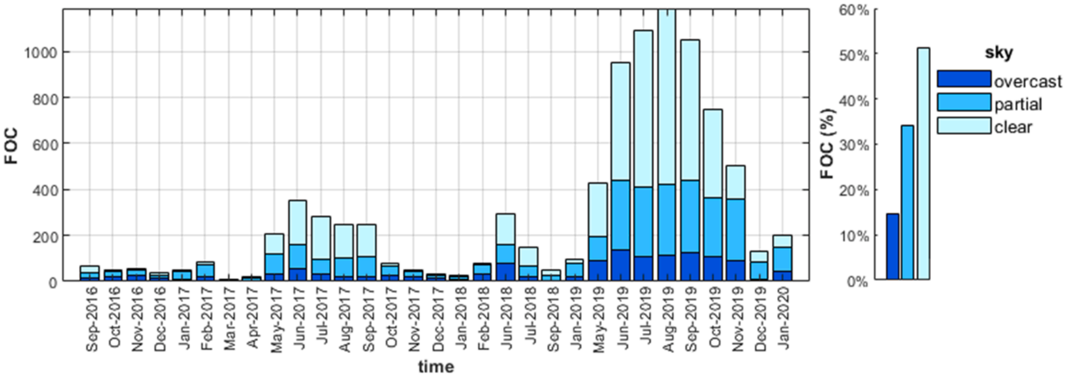

Fig. 3. Monthly distribution of the Frequency of Occurrence (FOC) and total FOC (%) of clear, partly cloudy, and overcast sky conditions in Burgos, Spain (from September 21st, 2016 to January 31st, 2020).

effectively contributing to better definition of the atmospheric conditions.

The cloud ratio on irradiance, C_e (Rahim et al., 2004) originally defined as the proportion of diffuse to global irradiance, was used in the estimation of solar radiation. Umemiya and Kanou (2008) introduced a new definition in terms of illumination, C_v, and used it for sky tree classification. The cloud ratio is 1 in overcast sky conditions and 0 for clear skies, and it will vary quickly and with some frequency when the sky is partly cloudy. The cloudless index, *Cle*, is often defined in terms of the standard cloud ratio, C_{es}, and the cloud ratio, C_e. C_{es} is defined as a polynomic fit of the lower limit of C_e.

Perraudeau's nebulosity index, *OFP*, introduced by Perraudeau in 1989 (Kambezidis et al., 1998), classifies the skies into five categories (Kambezidis, 2018). This index has since been modified by other authors (Fakra et al., 2011) and is defined in this work as *FP*. The clearness function, *F*, was compared to the MIs Δ, ε_p, and k_t for sky classification (Muneer, 2007). Low values of *F*, indicate overcast sky conditions and values near to 1 are obtained under clear sky conditions.

The Klucher index, *FK*, (Klucher, 1979) depending only on k_d, has also been used for sky classification. Markou et al. (2005) prepared a simple sky classification by modelling direct solar irradiance data, P_e, characteristic of each sky type. This proposal suggested the use of experimental MIs for sky classification: horizontal global irradiance, *RaGH*; horizontal diffuse irradiance, *RaDH*; horizontal beam irradiance, *RaBH*; and south-facing global vertical irradiance, *RaGVS*.

Umemiya and Kanou (2008) proposed the turbidity index, *TURV*, permeability, *PERM*, Unemiyas's Cloud Ratio, *CLDV*, and global, diffuse, and beam illuminance, *EVGM*, *EVDM*, and *EVSM*, normalized to the optical mass, M_v, as effective sky condition sorters. They produced a sky classification with 7 types of skies that used a classification tree based on the turbidity index, *TURV*; Kittler's index, k_t; sky brightness, Δ; and, normalized global illuminance, *EVGM*. A similar proposal was introduced by Lou et al. (2017) using solar altitude, α_s; Kittler's index, k_t; the turbidity index, *TURV*; air temperature, *T*; and, relative humidity, *RH*. Other variables, used for meteorological forecasting have been proposed among which MIs for sky classification such as wind speed, *WS*; relative humidity, *RH*; cloud cover, *CC*; and air temperature, *T*, among others (Inman et al., 2013).

Li *et al.* [12] proposed a group of MIs that obtained a very accurate sky classification. They used a ratio of zenith illuminance, L_z, and horizontal diffuse illuminance, *LxDH*, named *LERT*, as a measure of sky brightness (Li et al., 2006; Markou et al., 2005). The luminous turbidity index, t_v, refers to the attenuation of solar radiation in the atmosphere, due to the molecules contained into the air (water, dust or aerosols) (Li et al., 2016; Pasero and Mesi, 2010). In overcast sky conditions, t_v, is very high, because there is no direct solar-irradiation component. Under clear or partly cloudy sky conditions, t_v is a very interesting parameter, due to its high sensitivity to ambient pollution (Lou et al., 2017). It is related to CIE standard sky types VI.6, VI.5, and IV.4 (Kocifaj, 2011).

Table 4
Definition of the 43 MIs reviewed as candidates for sky classification. L_0 is the Luminous solar constant (133.8 kLux) and I_{SC} is the standard global irradiance ($1361.1\frac{\mathrm{W}}{\mathrm{m}^2}$) (Gueymard, 2018).

Ratio Zenith Illuminance to horizontal diffuse Illuminance	*Ratio Global Illuminance*	*Ratio Diffuse Illuminance*	*Luminous Turbidity index*	*Vertical Sky Component*
$LERT = \frac{L_Z}{LxDH}$	$C1 = \frac{LxGH}{L_{Oh}}$	$C2 = \frac{LxDH}{L_{Oh}}$	$t_v = \frac{Ln(\frac{L_{Oh}}{LxBH})}{A_V M_V}$	$VSC = \frac{RaDH}{RaDV}$
Normalized Global Illuminance $EVGM = M_V \frac{LxGH}{L_0}$	***Normalized Beam Illuminance*** $EVSM = M_V \frac{LxBH}{L_0}$	***Normalized Diffuse Illuminance*** $EVDM = M_V \frac{LxDH}{L_0}$	***Cloudless Index*** $Cle = \frac{1-k_d}{1-Ces(M)}$	***Igawa's Sky Index*** $S_i = \frac{RaGH}{0.84\frac{I_{SC}}{M_v}e^{-0.0675M_v}} + \sqrt{Cle}$
Direct Fraction $k_b = \frac{RaBH}{RaGH}$	***Cloud Cover*** $CC(\%Clouds)$	***Illuminance Cloud Ratio*** $C_V = \frac{LxDH}{LxDH + LxBH}$	***Irradiance Cloud Ratio*** $C_e = \frac{RaDH}{RaDH + RaBH}$	***Standard Cloud Ratio*** $Ces = 0.01299 + 0.07698M_V - 0.003857{M_V}^2 + 0.0001054{M_V}^3 - 0.000001031{M_V}^4$
Umemiya's Cloud Ratio $CLDV = \frac{LxDH}{LxGH}$	***Relative Heaviness*** $\Omega = \frac{LxGH}{Sin\alpha_s}$	***Clear Sky Index*** $k_C = \frac{LxGH}{0.84\frac{I_{SC}}{M_v}e^{-0.0675M_v}}$	***Clearness Index*** $k_t = \frac{RaGH}{I_O Sin\alpha_s}$	***Zenith Angle Independent Clearness Index*** $k_{t2} = \frac{k_T}{1.031e^{-\frac{1.4}{0.9+9.4M_V}} + 0.1}$
Luminous Efficacy $K = \frac{LxGH}{RaGH}$	***Brightness Index*** $\Delta = \frac{RaDHM_v}{I_{sc}\varepsilon_o \sin\alpha_s}$	***Perez's Clear sky index*** $\varepsilon_p = \frac{\frac{(RaDH + RaBH)}{RaDH} + 1.04Z^3}{1 + 1.04Z^3}$	***Original Perraudeau's Nebulosity Index*** $OFP = \frac{1-k_D}{1-\frac{E_{clear}}{E_{clear}} + RaGH}$	***Perraudeau's Nebulosity Index*** $FP = \frac{1-k_D}{1-0.12037(SinZ_S)^{-0.82}}$
Klucher's Clearness Index $FK = 1 - k_D^2$	***RaBH, RaDH, RaGHRaGVS, LxGH*** *Direct, Diffuse, Global (Horizontal and Vertical South oriented) Irradiance global horizontal Illuminance*	***Optical Mass*** $M_v = (sin\alpha_s + 0.50572(\alpha_s + 6.07995)^{-1.6364}$	***Scattering Angle Ref*** $\chi = arcos(cosZ_s cosZ_P + sinZ_s sinZ_P cos\lvert\phi_P - \phi_s\rvert)$	***Turbidity*** $TURV = \frac{1 + 0.0045M_V}{0.1M_V} Ln(\frac{L_0}{LxBH})$
T, RH, WS, L_z, α_s *Temperature, Relative humidity, Wind speed, Zenith luminance, solar altitude.*	***Diffuse fraction*** $k_d = \frac{RaDH}{RaGH}$	***Clearness Function*** $F = \frac{RaGH - RaBH}{I_{sc}\varepsilon_o \sin\alpha_s}$	***Modeled direct solar irradiance*** $P_e = \frac{RaGH - RaDH}{sin(\alpha_s)}$	***Permeability*** $PERV = M_V\sqrt{\frac{LxBH}{L_0}}$

$C1$, defined as the ratio of horizontal global illuminance, *LxGH*, and horizontal extraterrestrial illuminance, L_{0h}, evaluates the ambient clarity. Low values of $C1$ are characteristic of the passage of a cloud on a clear day while a high $C1$ value can reflect a cloud opening zone on a completely overcast day (Alshaibani, 2016a; Kittler and Danda, 2000). $C2$ is defined as the ratio of horizontal diffuse illuminance, LxDH, and horizontal extraterrestrial illuminance, L_{0h}, so high $C2$ values are characteristic of partly cloudy skies, while low $C2$ values are characteristic of cloudy or completely clear skies (Li et al., 2006; Li et al., 2010; Markou et al., 2005).

The vertical sky component, *VSC*, was also proposed as an MI for sky classification (Li et al., 2011b). Defined as the ratio of the vertical diffuse illuminance and horizontal diffuse illuminance, it can easily be obtained experimentally. Littlefair established an international standard for the indoor daylight evaluation of buildings (Littlefair, 2012) based on *VSC*, which is highly dependent on the solar altitude, α_s, and the scattering angle, χ (Alshaibani, 2011; Li et al., 2014a).

2.4. Feature selection

43 MIs were selected (Table 1) for the study. Each one represents certain characteristics of the sky that are suitable for sky classification. The final objective of the present work is to distinguish the most representative MIs for sky classification according to the CIE taxonomy.

The most simple and demanding methodology is the full combinatorial method. It proposes to test all the possible combinations of all MIs: at first, only one MI would be considered for the CIE classification; then, all combinations of two MIs would be used for the task and so on (Visa et al., 2011). Li et al. (2011a) followed this path to evaluate the performance of several MIs in neural networks for weather data classification. It was feasible because only five MIs were considered which meant a total of 30 sets of MIs for testing. In the present study, an analysis of the 43 variables implied over a trillion combinations, which was not feasible. FS was therefore essential to solve this task.

Several FS techniques are used widely in the ML field to find the most important variables or to reject the most redundant ones. There are different types of FS algorithms. On the one hand, no clustering algorithms are used with the Filter Methods that base their decision on a statistical index that evaluates the dependence between the MIs. On the other hand, the Wrapper Methods evaluate the information provided by each MI using clustering algorithms, which implies a higher computational cost (Solorio-Fernández et al., 2019).

The FS Filter methods are used to study the similarity of MIs through a statistical parameter, a mathematical expression that serves to eliminate redundant or non-informative indices. These methods are independent from the ML algorithm used later on (Mitra et al., 2002; Yu and Liu, 2003). Hence, their results may be used as an input of any ML algorithm. They are an efficient procedure, in so far as they reduce the input dimensionality of the ML algorithm and prevent overfitting. In this study, a widely used statistical parameter will be used: the Pearson correlation coefficient.

The FS Wrapper methods perform a global evaluation of the entire set of variables that creates a ranking of relevance. The ML algorithm executes the ranking and, consequently, the score is not universal and they cannot be applied to any other ML algorithm (Wald et al., 2014). In other words, the Wrapper methods will produce different rankings for different ML algorithms. Wrapper FS approaches are commonly used in the field of renewable energy applications due to their higher performance (Salcedo-Sanz et al., 2018). Permutation Importance, Recursive Feature Elimination and Boruta methods are all included within this category. In this work, the FS algorithms used for simplifying the classification trees are: Pearson correlation coefficient, Permutation

Importance, Recursive Feature Elimination and Boruta. The following paragraphs describe them and their use in other fields of the ML.

2.4.1. *Pearson correlation coefficient criterion (P)*

The Pearson criterion is based on the Pearson correlation coefficient, ***r***. If two datasets X and X' are strongly correlated, the Pearson coefficient is 1 (direct correlation) or −1 (inverse correlation). However, a Pearson coefficient near 0 implies a weak or null correlation.

In this work, the Pearson criterion was applied in two steps: firstly for selecting the MIs with a strong correlation to the CIE cloudiness classification. Only the MIs with Pearson correlation coefficients above a certain threshold were selected and used in the next step for detecting the MIs with high correlations between them and for selecting the most important ones. After both steps, only the most important independent MIs for the classification were selected.

2.4.1.1. Permutation Importance (PI). Permutation Importance (PI) or the Mean Decrease in Accuracy (MDA) (Nembrini, 2019) algorithm is used to analyze how the score of the prediction model decreases when the data of a single variable is randomly permuted, generating random noise. Permutation feature importance is defined as decreasing in a model score when a single feature value is randomly shuffled (Bommert et al., 2020). A PI index of 0% means null relevance of this feature for the classification. Usually, a threshold of 5% is employed, considering only MIs with a permutation importance above 5% as important and discarding any others (Altmann et al., 2010).

2.4.1.2. Recursive feature Elimination (RFE). The Recursive Feature Elimination (RFE) method fits a model, so as to remove the weakest features until a specified number of variables is reached. A great number of ML classification algorithms such as Decision Trees, Support Vector Machines (Weston et al., 2001), and Random Forests (Diaz-Uriarte and Alvarez de Andres, 2006), among others, attach a weight to each input for the classification. The features are ranked in each loop and a few features per loop are removed, in an attempt to lower their interdependencies and collinearity. Also, the final size of the feature set cannot be initially specified and the number is established when there is no global improvement in the accuracy of the model. This method has been widely used with high-dimensional data sets (Escanilla et al., 2018; Paul et al., 2015). Fields where the algorithm has successfully been applied include genetics (Darst et al., 2018), materials science (Sharp et al., 2018), cancer studies (Duan et al., 2005), sports (Paul et al., 2015) and solar and wind forecasting (Benamrou et al., 2020; Feng et al., 2017).

2.4.1.3. Boruta (BOR). The Boruta (BOR) method, rather than comparing features between each other, competes with a randomized version of so-called "shadow features". In each iteration, the importance given by the classification algorithm to each original feature is compared with the highest feature importance recorded among the shadow features. Each time the importance of a feature is higher than this threshold, it is called a "hit". A feature is considered useful, if it performs better than the best randomized feature. Counting the number of hits, the selection of a feature is decided after a number of trials. In the same way as RFE, the BOR method performs a top-down search for relevant features, progressively eliminating irrelevant ones (Kursa and Rudnicki, 2010). RFE and Boruta have been compared on many occasions and in scientific fields such as genetics (Kursa, 2014) and spectroscopy (Poona et al., 2016). Permutation Importance is highly sensitive and effective (Gregorutti et al., 2017) when applied to biological data (Degenhardt et al., 2019).

2.5. *Classification trees*

A classification tree is an algorithm that classifies datasets into certain outcome categories by using a sequence of "partitions", or "splits". However, the more complex the category analysis, the larger the sequence of splits that may be needed. Since the first implementation of Breiman *et al.* in 1984 (Breiman, 1984), classification trees have been used in a very large variety of disciplines, such as meteorology, medicine, and statistics, among others, and likewise CIE Standard skies classification (Umemiya and Kanou, 2008).

The structure of the classification tree can be implemented by several criteria. The one chosen for the sky cloudiness classification is the Classification and Regression Tree (CART) (Breiman, 1984). It looks for successive binary splits that chooses the partitions, in order to obtain the highest performance. Both the Gini (D'Ambrosio and Tutore, 2011) and the Entropy (Witten et al., 2016) indices were considered to fit the classification tree. The Gini Index points to how often a randomly chosen element from the set would be incorrectly labelled. The entropy index considers the disorder of a grouping by the target variable. Both of them are performance measures of the classification tree.

The classification tree algorithm was selected above other ML classification algorithms, due to the transparency of the results it can obtain. The classification tree produces a diagram that can be more easily understood than those produced by other ML techniques such as, Support Vector Machines, Neural Networks, Random Forest and Gradient Boosting, traditionally known as "black boxes".

2.6. *Classification metrics*

Confusion matrices are a useful tool for the performance characterization of an classification algorithm. Four possible cases can be obtained in a classification procedure: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The number of each one in the confusion matrix summarizes the performance of a dichotomic classification, as shown in Fig. 4.

From the confusion matrix Precision, *Pr*, and Recall, *Re*, indices are defined in Eqs. (1) and (2), respectively, in order to measure the performance of the classification algorithm. *Pr* is the probability that a positive prediction was correct, while *Re* is the percentage of correctly detected positive cases. Both indices are independent from each other and could be explained by a very precise and not a very sensitive algorithm. Both indices are grouped in the $f1$ factor, defined in Eq. (3) as the harmonic average of precision and recall.

$$Pr = \frac{TP}{TP + FP} \tag{1}$$

$$Re = \frac{TP}{TP + FN} \tag{2}$$

$$f1 = \frac{2}{\frac{1}{Pr} + \frac{1}{Re}} = \frac{2TP}{2TP + FN + FP} \tag{3}$$

Sky classification is a multiclass classification. It is therefore necessary to sum up the indices for each category, which yields a global result for the classification algorithm performance, as shown in Fig. 5. *Pr*, *Re*, and *F*1 indices were calculated for each CIE Standard sky condition (Clear, Partial and Overcast).

Two different procedures can be used to obtain the global values: the macro-average and the weighted-average. The macro-average calculates the global value for each index as the average of the index for each category, regardless of the size of the category within the sample. Therefore, a low performance of the classification algorithm in one of the categories may affect overall performance, despite performing well in the other categories. This problem is as common in imbalanced class distribution as it is for the case of sky classification (see Fig. 2, where the FOC of clear, partial, and overcast conditions differ). The weighted-average yields the global value, by adding the results for each category and the weighting that represents the category size over the total number of cases. Therefore, the weighted aggregation used in this work

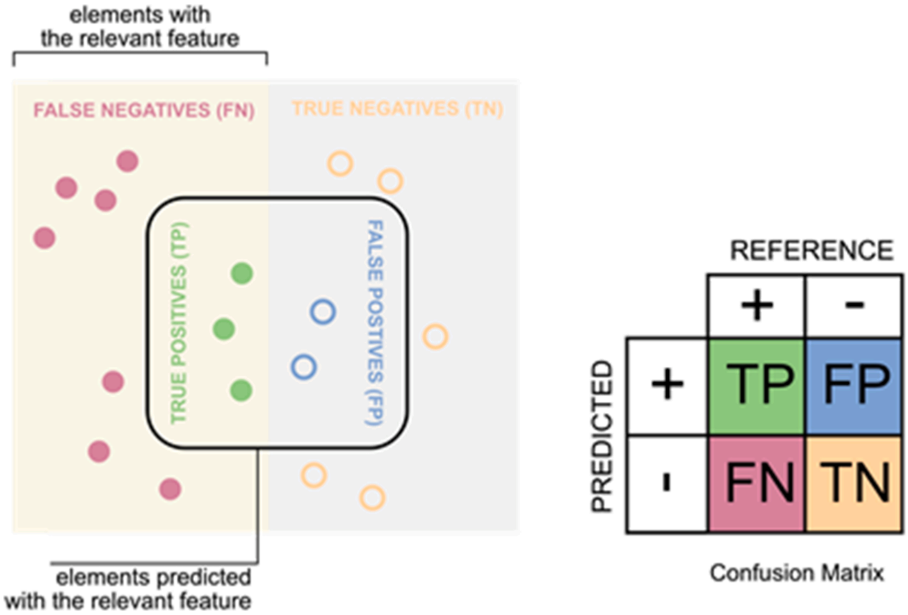

Fig. 4. Confusion matrix: possible cases in the comparison of the prediction with the actual data.

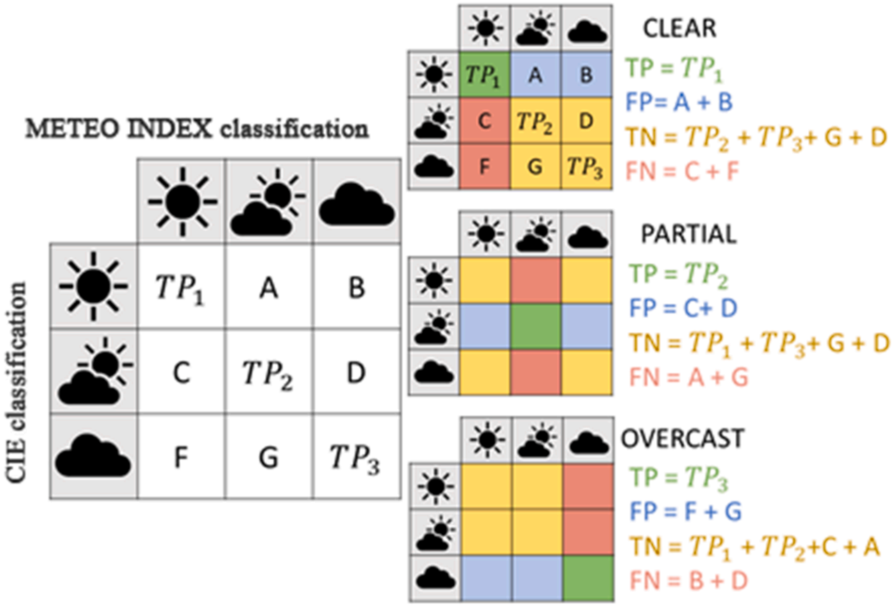

Fig. 5. Confusion Matrix for multi-class sky classification.

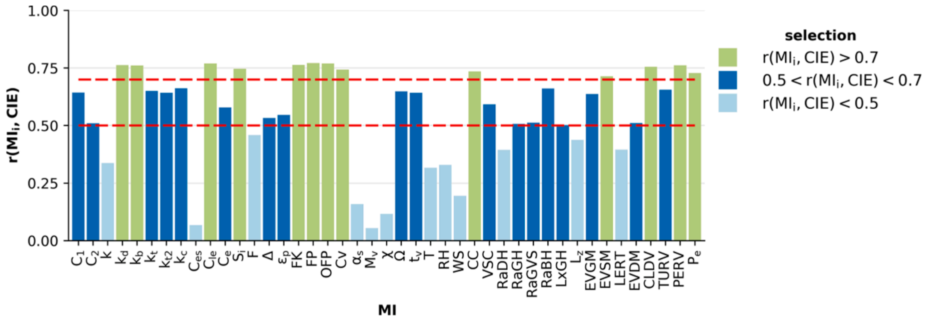

Fig. 6. Absolute value of the Pearson correlation between the MIs under consideration and the CIE Standard Sky classification, $r(MI_i, CIE)$.

will reward those algorithms with good performance in the most numerous classes and will have a lesser effect on those with poor performance, in the classes with fewer samples.

3. Results

3.1. Feature selection

Figs. 6 and 7 show the results of the FS using the Pearson correlation coefficient. Fig. 6 shows the absolute value of the correlation between each MI and the CIE classification, denoted as $r(MI_i, CIE)$. Following Thumb's rule (Mukaka, 2012), three r intervals were considered: high ($0.9 \geq r \geq 0.7$), moderate ($0.7 \geq r \geq 0.5$), and negligible ($r < 0.5$) correlations. The MIs with $r(MI_i, CIE) \geq 0.7$ were moved to the second stage of the Pearson correlation coefficient criterion. In the second step, an effort was made to discard the redundant MIs. Here, two MIs, $\mathrm{MI_i}$ and $\mathrm{MI_k}$, are redundant if $r(MI_i, MI_k) > 0.9$. Fig. 7 represents the correlation between the MIs. The MIs for which $r(MI_i, MI_k) \geq 0.9$. are shaded in blue. The redundant groups of the MIs were formed by grouping the MIs with very high correlation represented as blue squares, for each column of the matrix of Fig. 7. Each MI was only included in one group, as shown in Table 5. In this Table 5, all MIs in the same group were considered to have the same information and only one of them, the one with the highest value of $r(MI_i, CIE)$, was needed to reflect the information of the rest. *CC* was related to the others MIs and was therefore included in the selection process. As can be seen, the original set of 43 MIs was reduced to two. 13 features (MIs) were selected from the Permutation Index (PI) results as necessary for CIE Standard Sky classification, as shown in Fig. 8. All of them caused a 5% decrease in the performance of the classification algorithm when they were randomly shuffled. The red line in the figure represents the aforementioned threshold. All the scores above the line, represent an impact higher than 5%. Other thresholds could be considered. However, the optimal threshold for each algorithm is a matter for further research.

Thirteen MIs were selected using the Recursive Feature Elimination (RFE) FS procedure that sets a minimum number of MIs needed for accurate CIE Standard Sky classification at 13 MIs. Fig. 9 shows the curve produced by recursive reduction of the number of MIs. The curve maintains an excellent $f1$ above 13 features when the most suitable variables that the algorithm selected were CC,$LxGH$, VSC, WS, OFP, C_{le}, k_t, K, $LERT$, C_e, RH, L_z. When fewer variables were in the classification tree, the performance of the classification algorithm drastically decreased, because the most informative features were removed from the model. Conversely, redundant information was included, whenever additional variables were added.

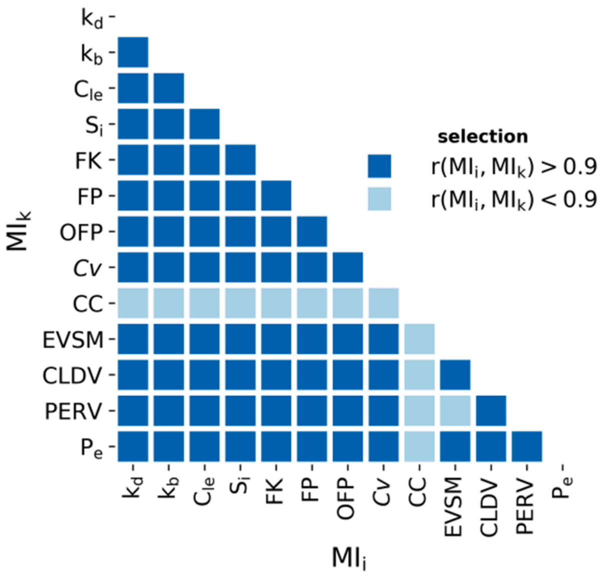

Fig. 7. Pearson correlation absolute value between MIs, $r(MI_i, MI_k)$, for MIs with $r(MI_i, CIE) \geq 0.7$.

One hundred trial tests of the Boruta (BOR) FS methodology were completed. Fig. 10 shows the hits of each one of the MI. All MIs got a hit and the maximum number of hits was below ten. The MI with a number of hits higher than one was selected.

The MIs selected by each FS procedure are summarized in Table 6. With the exception of the Boruta method, the FS procedures reduced the original set of MIs to a little less than 75%, selecting different MIs. The reduction in the number of variables required for the classification process, reflects the usefulness of the FS. Fewer variables to be measured and/or calculated implies less instrumentation and data storage, and simplifies the classification algorithm. Simpler models reduced the necessary computing power and, for example, made its implementation easier for lighting control systems.

The results of different feature methods selection, show the relationship existing between the variables, which in some cases can be directly deduced from the definition thereof, shown in Table 1 while in other cases does not appear so clearly reflected. The MIs selected by the FS algorithms can be classified into three types: variables related to the cloud conditions, others related to daylighting, brightness or clearness conditions of the skies and geometrical variables. While the Pearson FS method eliminates those variables that are most related to each other, in order not to include redundant information, the Boruta method does not eliminate a priori, highly related variables that can add distinctive nuances useful for classification. PI and RFE methods reach a compromise between information and complexity.

3.2. Classification trees

The classification trees for CIE Standard Sky Classification from the MIs selected by Pearson, Permutation Importance, RFE and Boruta FS procedures are shown in Figs. 11–14. Starting in the main left node, if the condition is met, the path of the upper branch is followed and, if not, the path followed is the one indicated by the lower branch. Evaluating each node consecutively, the sky conditions would be obtained. The number inside the nodes represents the number of samples inside each partition. The number of binary partitions or levels of the classification tree is a previously set parameter. In this work all the classification trees have four levels. An increased number of levels might increase the precision of the classification algorithm in the same way as complexity. The starting MI and the number of levels of the classification tree were selected following the Gini and the Entropy criteria, previously introduced.

FP and *CC* are MIs selected by the Pearson FS method for the classification tree. Both MIs are related to the cloud conditions, through the diffuse horizontal fraction (ratio diffuse horizontal irradiation to global horizontal irradiation) and the percentage of sky covered by clouds, respectively. The CIE Standard decision tree obtained from the variables selected by the Pearson FS method identifies the clear sky type by one of these cases: $a)FP > 0.51$, and $CC \geq 0.53; b)FP > 0.78; c)FP \leq 0.51$and $CC \leq 0.66$.

Although the PI FS methods selected 11 MIs for the CIE sky classification, only three were used in the four-level classification tree: FP, CC, and VSC. The Vertical Sky Component, VSC, linked the classification to the daylighting.

The classification tree obtained with the MIs selected by the RFE FS method started with the CC, a variable which directly classified the skies as clear if <61.7%. On the second and third levels of the classification tree, original Perraudeau's Index, OFP, and VSC, were evaluated. At the last level, the MI selected to fit the classification was luminous efficacy, K. Again, two of the MIs were related to daylighting (VSC, K) and OFP and CC were related to cloud coverage.

Boruta FS methods selected 34 MIs for the CIE standard sky classi-

Table 5
Results of Pearson FS method.

Group 1	k_d	k_b	*Cle*	*Si*	*FK*	*FP*	*EVSM*	*CLDV*	*PERV*	P_e	*OFP*	C_V
$r(MI_i, \mathrm{CIE})$	0.762	0.761	0.770	0.747	0.764	0.771	0.714	0.755	0.762	0.729	0.769	0.740

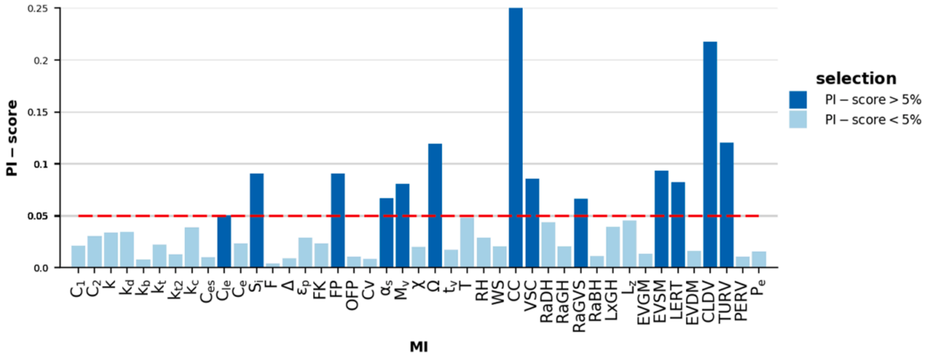

Fig. 8. Permutation Index (PI) results in feature selection of MIs for CIE Standard Sky classification.

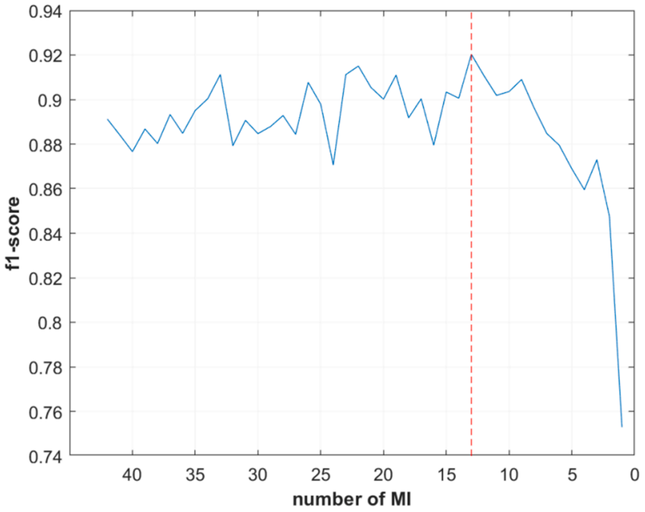

Fig. 9. Recursive Feature Elimination (RFE) results in FS of MIs for CIE Standard Sky classification.

fication, but four were necessary to build the four-level classification tree. The sky classification started by evaluating Perraudeau's nebulosity index, *FP* (Kambezidis et al., 1998). At the second level, cloud cover, *CC*, and the vertical sky component, *VSC*were investigated. Finally, the scattering angle, χ, a geometrical variable, was investigated.

As regards the intervals established by the classification trees for each partition with respect to the one established by the authors in their original works, it is important to remark that the number of sky categories is different for some MIs (for example *FP* establishes 5 sky categories, instead of three). However, the original intervals and those obtained were in consonance.

3.3. Analysis of the classification trees using different metrics

Fig. 15 shows the results of the *Pr*, *Re*, and *f*1 metrics obtained for the classification trees calculated from the MIs selected by each FS

Table 6
Summary of the features (MIs) selected by each one of the FS algorithms.

Feature Selection	MI selected	Number
Pearson correlation coefficient criterion (P)	FP, CC	2
Permutation Importance (PI)	$C_{le}, S_i, FP, \alpha_s, M_v, \Omega, CC, VSC, RaGVS, EVSM, LERT, CLDV, TURV$	13
Recursive Feature Elimination (RFE)	$K, k_t, C_{le}, C_e, OFP, \chi, RH, WS, CC, VSC, LxGH, L_z, LERT$	13
Boruta (BOR)	$C_1, C_2, CC, CLDV, C_e, \chi, Cle, C_v, \varepsilon_p, EVDM, EVGM, EVSMF, FK, FP, LERT, L_z, OFP, \Omega, PERV, P_e, RaBH, RaDHRaGVS, \Delta, \mathrm{S_i}, TURV, VSC, k_b, k_c, k_d, k_t, k_{t2}, t_v$	34

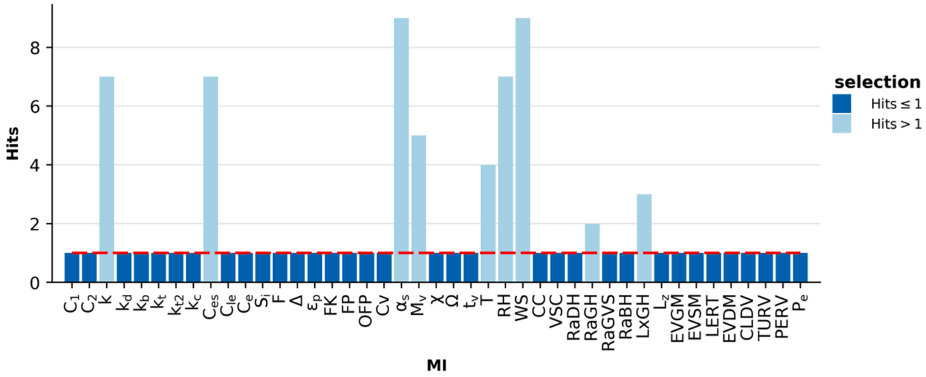

Fig. 10. Results of the Boruta FS methodology for CIE Standard Sky Classification after 100 trial tests.

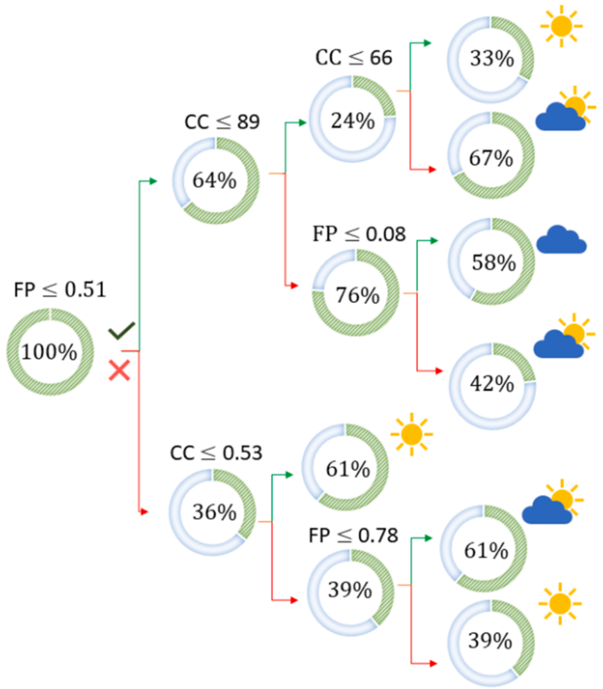

Fig. 11. CIE standard sky classification tree (MIs selected with the Pearson FS method).

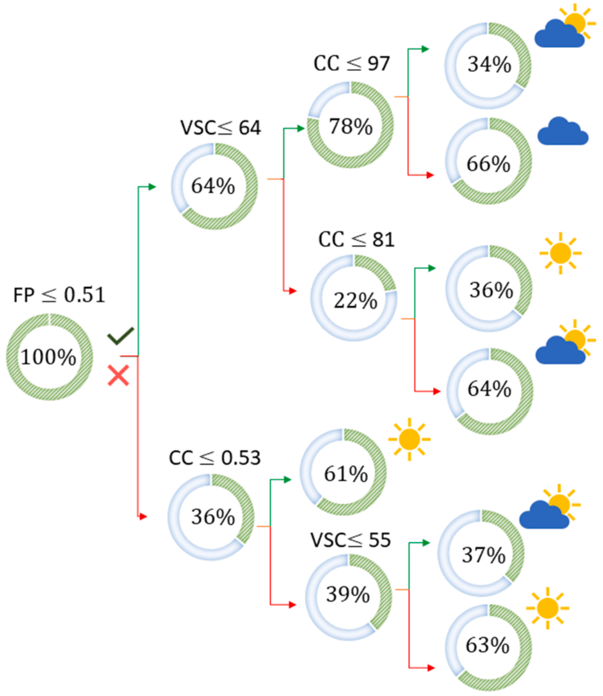

Fig. 12. CIE standard sky classification tree (MIs selected with the PI FS method).

procedure and for the different sky conditions (clear, partial, and overcast). The best results for all metrics were obtained in overcast sky conditions, with *Pr* and *Re*above 85% and reaching 90% in the case of the classification tree that applied the four MIs selected for the RFE method (CC, OFP, VSC, K). The classification trees that used the MIs selected by both PI (FP, VSC, CC), and BORUTA (FP, VSC, CC, χ) obtained the same results. The simplest classification tree, from the two MIs selected by the P method, (CC, FP), yielded worse $f1$ metrics.

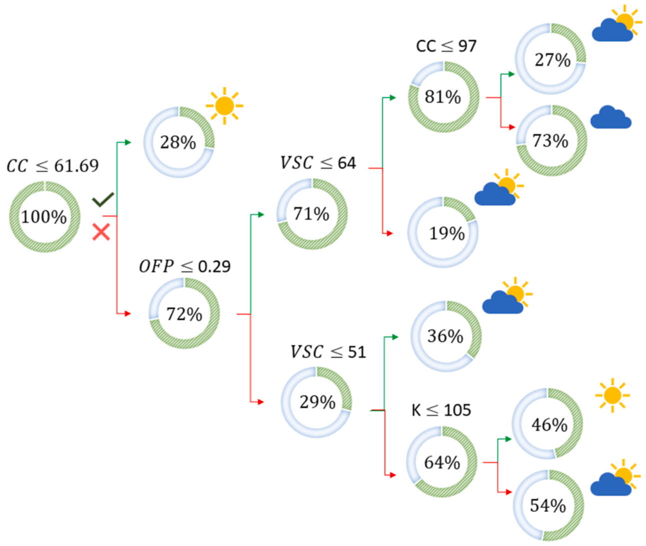

Fig. 13. CIE standard sky classification tree (MIs selected with the RFE FS method).

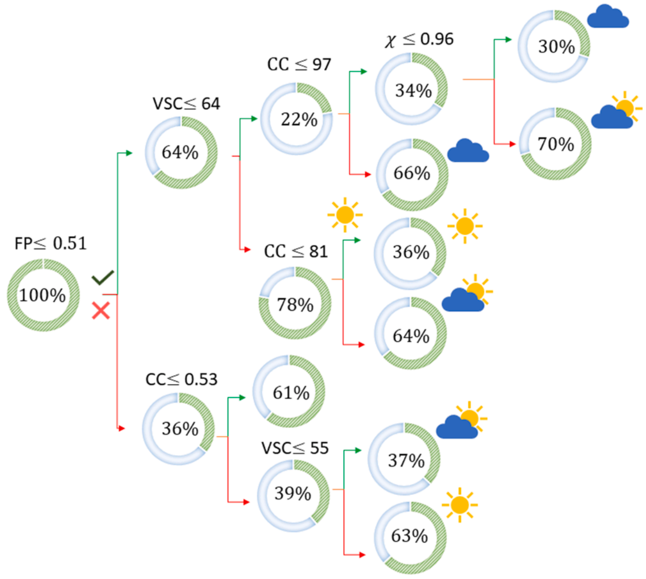

Fig. 14. CIE standard sky classification tree. (MIs selected with the Boruta FS method).

In the identification of clear sky conditions, all classification trees presented more dispersion of the metrics value: $f1$ ranged from 65% (P method) to 87% (BORUTA method), but all indices exceeded 65%. The identification of partially covered skies was worse, lowering the values of all indices by between 55% and 70%. The irregularity of partially covered sky conditions, the high variability of the MIs for these conditions and the dependency of cloud cover with respect to the Sun might explain this fact. In every case, the RFE FS method yielded the closest values of the three metrics.

The weighted-averaged global $f1$ is shown in Fig. 16. As can be seen, all the classification trees yielded results between 74% and 77%, the highest value of which was produced by the RFE FS procedure, very close to the BOR and PI FS procedure with the same value of $f1$. The Pearson method, also the simplest classification tree, showed the lowest $f1$ value. Taking into account the number of MIs used by each classification tree, perhaps the RFE FS methods offered the best performance with no high complexity, but the results highlighted no significant

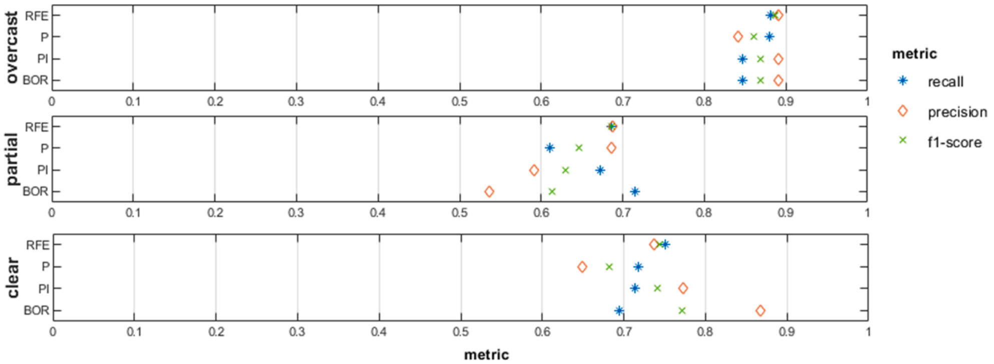

Fig. 15. Precision, Pr, Recall, Re and $f1$ indices calculated for each CIE Standard sky type for each classification tree based on the different FS procedures.

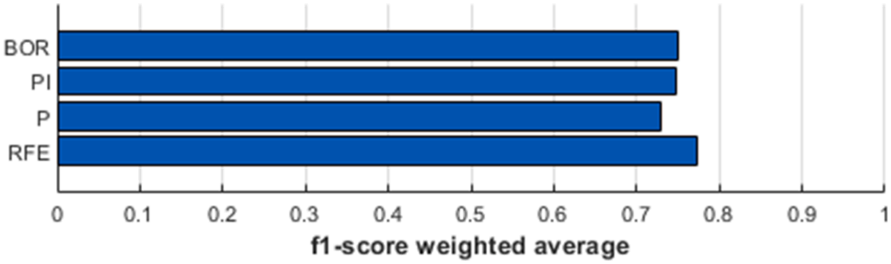

Fig. 16. Weighted-averaged global $f1$.

advantages between the classification algorithms constructed from the different feature selection procedures used in this work.

4. Conclusions

This study has highlighted the usefulness of the FS procedure for adequate determination of MIs for sky classification in accordance with the CIE Standard classification, as an alternative to the use of sky-scanner devices. The maximum number of MIs can be identified with FS for use as an input for the ML algorithm, avoiding the introduction of redundant and useless information. Four FS (filter and wrapper) methods have been reviewed and applied. The initial set of 43 MIs was drastically reduced by three of the FS algorithms (Pearson, PI and RFE), although a less significant reduction was achieved with the Boruta FS method. The main advantage of the Pearson FS procedure over and above all the other methods that were tested was its independence from the ML algorithm used after the FS procedure, with the consequent saving of time when it was necessary to verify the operation of different ML algorithms.

All the classification trees yielded performances that were similar to the CIE standard sky classification in terms of the Pr, Re and $f1$ metrics. The worse results were shown for the identification of partially cloudy conditions, while the overcast and clear sky conditions were identified with high success rates. No significant differences in the performance of the classification algorithms constructed from the MIs selected by the different FS methods have been pointed out, and the use of one or another FS method could be at the discretion of the researcher.

The MIs selected by the FS algorithms can be classified into three types: variables related to the cloud conditions, (FP, CC, OFP), others related to daylighting, brightness or clearness conditions of the skies (VSC, K), and geometrical variables, such as α_s and χ.

Both the intervals established by the classification trees for each partition and those established by the authors in their original works were in consonance. However, the classification tree might be a good alternative, in order to set up these intervals independently from local climatic and meteorological conditions.

Supplementary Materials

Experimental data and Phyton code in this paper are available with copyleft licenses for the data and the code in the following link: http://hdl.handle.net/10259/5563 (DOI: 10.36443/10259/5563).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge the financial support provided by the Regional Government of Castilla y León under the “Support Program for Recognized Research Groups of Public Universities of Castilla y León” (BU021G19) and the Spanish Ministry of Science & Innovation under the I + D + i state program “Challenges Research Projects” (Ref. RTI2018-098900-B-I00). Diego Granados López express his thanks to Junta de Castilla-León for economic support (PIRTU Program, ORDEN EDU/556/2019).

References

Alshaibani, K., 2011. Finding frequency distributions of CIE Standard General Skies from sky illuminance or irradiance. Light. Res. Technol. 43 (4), 487–495. https://doi.org/10.1177/1477153511404999.

Alshaibani, K., 2016a. Average daylight factor for the ISO/CIE Standard General Sky. Light. Res. Technol. 48 (6), 742–754. https://doi.org/10.1177/1477153515572939.

Alshaibani, K., 2016b. The use of horizontal sky illuminance to classify the CIE Standard General Skies. Light. Res. Technol. 48 (8), 1034–1041. https://doi.org/10.1177/1477153515624485.

Alshaibani, K., 2017. Classification Standard Skies: the use of horizontal sky illuminance. Renew. Sust. Energ. Rev. 73, 387–392. https://doi.org/10.1016/j.rser.2017.01.116.

Altmann, A., Tolosi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinform. 26 (10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134.

Allard, D., Ailliot, P., Monbet, V., Naveau, P., 2015. Stochastic weather generators: an overview of weather type models. Journal de la Société Française de Statistique 156 (1), 101–113. https://hal.inrae.fr/hal-02641587.

Aries, M.B.C., Aarts, M.P.J., Van Hoof, J., 2015. Daylight and health: a review of the evidence and consequences for the built environment. Light. Res. Technol. 47 (1), 6–27. https://doi.org/10.1177/1477153513509258.

Benamrou, B., Ouardouz, M., Allaouzi, I., Ben Ahmed, M., 2020. A proposed model to forecast hourly global solar irradiation based on satellite derived data, deep learning and machine learning approaches. J. Ecol. Eng. 21 (4).

Biesiada, J., Duch, W., 2007. Feature selection for high-dimensional data — a pearson redundancy based filter. In: Kurzynski M., P.E., Wozniak M., Zolnierek A. (Ed.) Computer Recognition Systems 2. Advances in Soft Computing. Springer, Berlin, Heidelberg, pp. 242–249.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M., 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. Comput. Stat. Data An. 143, 106839 https://doi.org/10.1016/j.csda.2019.106839.

Breiman, L., 1984. Classification and Regression Trees. Routledge, New York.

Comission Internationale de, L.E., 1995. Guide to recommended practice of daylight measurement, CIE 108-1994. Color Res. Appl. 20(1), 80–80. https://doi.org/10.1002/col.5080200118.

Chaiwiwatworakul, P., Chirarattananon, S., 2004. Distribution of sky luminance in tropical climate. In: Proceedings of the Joint International Conference on Sustainable Energy and Environment, Thailand, 1–3 December. pp. 530–537.

Chen, W., Li, D.H.W., Lou, S., 2019. Estimation of irregular obstructed vertical sky components under various CIE skies. Energy Procedia 158, 309–314. https://doi.org/10.1016/j.egypro.2019.01.094.

Chung, T.M., 1992. A study of luminous efficacy of daylight in Hong Kong. Energy Build. 19 (1), 45–50. https://doi.org/10.1016/0378-7788(92)90034-E.

D'Ambrosio, A., Tutore, V.A., 2011. Conditional classification trees by weighting the gini impurity measure. In: 7th Biannual Meeting of the Classification and Data Analysis Group, CLADAG 2009. Catania, pp. 273–280. https://doi.org/10.1007/978-3-642-11363-5_31.

Darst, B.F., Malecki, K.C., Engelman, C.D., 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet. 19 (Suppl 1), 65. https://doi.org/10.1186/s12863-018-0633-8.

Darula, S., Kittler, R., 2002. CIE general sky standard defining luminance distributions. Proc. eSim 11–13.

Darula, S., Kittler, R., Kómar, L., 2013. Sky type determination using vertical illuminance. Przegląd Elektrotechniczny 89 (6), 315–319.

Degenhardt, F., Seifert, S., Szymczak, S., 2019. Evaluation of variable selection methods for random forests and omics data sets. Brief. Bioinform. 20 (2), 492–503. https://doi.org/10.1093/bib/bbx124.

Diaz-Uriarte, R., Alvarez de Andres, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinform. 7, 3. https://doi.org/10.1186/1471-2105-7-3.

Dieste-Velasco, M.I., Díez-Mediavilla, M., Granados-López, D., González-Peña, D., Alonso-Tristán, C., 2019. Performance of global luminous efficacy models and proposal of a new model for daylighting in Burgos, Spain. Renew. Energy 133, 1000–1010. https://doi.org/10.1016/j.renene.2018.10.085.

Duan, K.-B., Rajapakse, J.C., Wang, H., Azuaje, F., 2005. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans. Nanobiosci. 4 (3), 228–234.

Dubois, M.C., Gentile, N., Amorim, C.N.D., Osterhaus, W., Stoffer, S., Jakobiak, R., Geisler-Moroder, D., Matusiak, B., Onarheim, F.M., Tetri, E., 2016. Performance evaluation of lighting and daylighting retrofits: results from IEA SHC task 50. Energy Procedia 91, 926–937. https://doi.org/10.1016/j.egypro.2016.06.259.

Edwards, L., Torcellini, P., 2002. Literature review of the effects of natural light on building occupants. National Renewable Energy Lab., Golden, CO.(US).

Escanilla, N.S., Hellerstein, L., Kleiman, R., Kuang, Z., Shull, J.D., Page, D., 2018. Recursive Feature Elimination by Sensitivity Testing. In: Proceedings of the … International Conference on Machine Learning and Applications. International Conference on Machine Learning and Applications 2018, pp. 40–47. https://doi.org/10.1109/ICMLA.2018.00014.

Fakra, A.H., Boyer, H., Miranville, F., Bigot, D., 2011. A simple evaluation of global and diffuse luminous efficacy for all sky conditions in tropical and humid climate. Renew. Energy 36 (1), 298–306. https://doi.org/10.1016/j.renene.2010.06.042.

Feng, C., Cui, M., Hodge, B.-M., Zhang, J., 2017. A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. Appl. Energy 190, 1245–1257.

Ferraro, V., Igawa, N., Marinelli, V., 2010. INLUX-DBR – A calculation code to calculate indoor natural illuminance inside buildings under various sky conditions. Energy 35 (9), 3722–3730. https://doi.org/10.1016/j.energy.2010.05.021.

Fouquart, Y., Buriez, J.C., Herman, M., Kandel, R.S., 1990. The influence of clouds on radiation: a climate-modeling perspective. Rev. Geophys. 28 (2), 145–166. https://doi.org/10.1029/RG028i002p00145.

Granados-López, D., Díez-Mediavilla, M., Dieste-Velasco, M.I., Suárez-García, A., Alonso-Tristán, C., 2020. Evaluation of the vertical sky component without obstructions for daylighting in Burgos, Spain. Appl. Sci. 10 (9), 3095.

Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and variable importance in random forests. Stat. and Comput. 27 (3), 659–678. https://doi.org/10.1007/s11222-016-9646-1.

Gueymard, C.A., 2018. A reevaluation of the solar constant based on a 42-year total solar irradiance time series and a reconciliation of spaceborne observations. Sol. Energy 168, 2–9. https://doi.org/10.1016/j.solener.2018.04.001.

Gueymard, C.A., Ruiz-Arias, J.A., 2016. Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. Sol. Energy 128, 1–30. https://doi.org/10.1016/j.solener.2015.10.010.

Igawa, N., Koga, Y., Matsuzawa, T., Nakamura, H., 2004. Models of sky radiance distribution and sky luminance distribution. Sol. Energy 77 (2), 137–157. https://doi.org/10.1016/j.solener.2004.04.016.

Inman, R.H., Pedro, H.T.C., Coimbra, C.F.M., 2013. Solar forecasting methods for renewable energy integration. Prog. Energy Combust. Sci. 39 (6), 535–576. https://doi.org/10.1016/j.pecs.2013.06.002.

ISO, 2004. ISO-15469:2004 (E). Spatial distribution of daylight-CIE standard general sky. Geneve, Switzerland.

Janjai, S., Tohsing, K., Nunez, M., Laksanaboonsong, J., 2008. A technique for mapping global illuminance from satellite data. Sol. Energy 82 (6), 543–555. https://doi.org/10.1016/j.solener.2007.11.003.

Kambezidis, H.D., 2018. The solar radiation climate of Athens: variations and tendencies in the period 1992–2017, the brightening era. Sol. Energy 173, 328–347. https://doi.org/10.1016/j.solener.2018.07.076.

Kambezidis, H.D., Muneer, T., Tzortzis, M., Arvanitaki, S., 1998. Global and diffuse horizontal solar illuminance: month-hour distribution for Athens, Greece in 1992. Light. Res. Technol. 30 (2), 69–74. https://doi.org/10.1177/096032719803000203.

Kittler, R., Danda, S., 2000. Determination of sky types from global illuminance. Int. J. Light. Res. Technol. 32 (4), 187–193.

Klucher, T.M., 1979. Evaluation of models to predict insolation on tilted surfaces. Sol. Energy 23 (2), 111–114.

Kocifaj, M., 2011. CIE standard sky model with reduced number of scaling parameters. Sol. Energy 85 (3), 553–559. https://doi.org/10.1016/j.solener.2010.12.024.

Kursa, M.B., 2014. Robustness of Random Forest-based gene selection methods. BMC Bioinform. 15 (1), 8.

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta package. J. Stat. Softw. 36 (11), 1–13.

Li, D., Lam, T., Wu, T., 2014a. Estimation of average daylight factor under obstructed CIE Standard General Skies. Light. Res. Technol. 46 (2), 187–197. https://doi.org/10.1177/1477153512453578.

Li, D.H.W., 2010. A review of daylight illuminance determinations and energy implications. Appl. Energy 87 (7), 2109–2118. https://doi.org/10.1016/j.apenergy.2010.03.004.

Li, D.H.W., Chau, T.C., Wan, K.K.W., 2014b. A review of the CIE general sky classification approaches. Renew. Sust. Energ. Rev. 31, 563–574. https://doi.org/10.1016/j.rser.2013.12.018.

Li, D.H.W., Cheung, G.H.W., Cheung, K.L., 2006. Evaluation of simplified procedure for indoor daylight illuminance determination against data in scale model measurements. Indoor Built Environ. 15 (3), 213–223. https://doi.org/10.1177/1420326x06066300.

Li, D.H.W., Cheung, K.L., Tang, H.L., Cheng, C.C.K., 2011a. Identifying CIE standard skies using vertical sky component. J. Atmos. Sol.-Terres. Phys. 73 (13), 1861–1867. https://doi.org/10.1016/j.jastp.2011.04.015.

Li, D.H.W., Lau, C.C.S., Lam, J.C., 2004. Standard skies classification using common climatic parameters. J. Sol. Energy Eng. 126 (3), 957–964. https://doi.org/10.1115/1.1740776.

Li, D.H.W., Lou, S., Lam, J.C., Wu, R.H.T., 2016. Determining solar irradiance on inclined planes from classified CIE (International Commission on Illumination) standard skies. Energy 101, 462–470. https://doi.org/10.1016/j.energy.2016.02.054.

Li, D.H.W., Lou, S.W., Lam, J.C., 2015. An analysis of global, direct and diffuse solar radiation. Energy Procedia 75, 388–393. https://doi.org/10.1016/j.egypro.2015.07.399.

Li, D.H.W., Tang, H.L., 2008. Standard skies classification in Hong Kong. J. Atmos. Sol.-Terr. Phys. 70 (8), 1222–1230. https://doi.org/10.1016/j.jastp.2008.03.004.

Li, D.H.W., Tang, H.L., Cheung, K.L., Lee, E.W.M., Cheng, C.C.K., 2011b. Sensitivity analysis of climatic parameters for sky classification. Theor. Appl. Climatol. 105 (3–4), 297–309. https://doi.org/10.1007/s00704-010-0392-6.

Li, D.H.W., Tang, H.L., Lee, E.W.M., Muneer, T., 2010. Classification of CIE standard skies using probabilistic neural networks. Int. J. Climatol. 30, 305–315. https://doi.org/10.1002/joc.1891.

Littlefair, P.J., 2012. Building Research Establishment, Site layout planning for daylight.

Lou, S., Li, D.H.W., Lam, J.C., 2017. CIE Standard Sky classification by accessible climatic indices. Renew. Energy 113, 347–356. https://doi.org/10.1016/j.renene.2017.06.013.

Lou, S., Li, D.H.W., Chen, W., 2019. A study of overcast, partly cloudy and clear skies by global illuminance and its variation features. IOP Conf. Ser.: Mater. Sci. Eng. 556, 012015. https://doi.org/10.1088/1757-899x/556/1/012015.

Markou, M.T., Kambezidis, H.D., Bartzokas, A., Katsoulis, B.D., Muneer, T., 2005. Sky type classification in Central England during winter. Energy 30 (9), 1667–1674. https://doi.org/10.1016/j.energy.2004.05.002.

Markou, M.T., Kambezidis, H.D., Katsoulis, B.D., Muneer, T., Bartzokas, A., 2004. Sky type classification in South England during the winter period. Build Res. J. 52 (1), 19–30.

Mitra, P., Murthy, C., Pal, S.K., 2002. Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intell. 24 (3), 301–312.

Mukaka, M., 2012. Statistics Corner: a guide to appropriate use of Correlation coefficient in medical research Malawi. Med. J.

Muneer, T., 2007. Solar Radiation and Daylight Models. Routledge, New York.

Muneer, T., Younes, S., Munawwar, S., 2007. Discourses on solar radiation modeling. Renew. Sust. Energ. Rev. 11 (4), 551–602. https://doi.org/10.1016/j.rser.2005.05.006.

Nembrini, S., 2019. On the behaviour of permutation-based variable importance measures in random forest clustering. J. Chemomet. 33 (8) https://doi.org/10.1002/cem.3135.

Ng, E., Cheng, V., Gadi, A., Mu, J., Lee, M., Gadi, A., 2007. Defining standard skies for Hong Kong. Handbook of Environmental Chemistry, Volume 5: Water Pollution 42 (2), 866–876. https://doi.org/10.1016/j.buildenv.2005.10.005.

Pasero, E., Mesi, L., 2010. Artificial Neural Networks for Pollution Forecast, Air Pollution. IntechOpen.

Paul, J., D'Ambrosio, R., Dupont, P., 2015. Kernel methods for heterogeneous feature selection. Neurocomputing 169, 187–195. https://doi.org/10.1016/j.neucom.2014.12.098.

Perez, R., Ineichen, P., Seals, R., Michalsky, J., Stewart, R., 1990a. Modeling daylight availability and irradiance components from direct and global irradiance. Sol. Energy 44 (5), 271–289. https://doi.org/10.1016/0038-092X(90)90055-H.

Perez, R., Ineichen, P., Seals, R., Zelenka, A., 1990b. Making full use of the clearness index for parameterizing hourly insolation conditions. Sol. Energy 45 (2), 111–114. https://doi.org/10.1016/0038-092X(90)90036-C.

Poona, N.K., van Niekerk, A., Nadel, R.L., Ismail, R., 2016. Random Forest (RF) wrappers for waveband selection and classification of hyperspectral data. Appl. Spectrosc. 70 (2), 322–333. https://doi.org/10.1177/0003702815620545.

Rahim, R., Baharuddin, Mulyadi, R., 2004. Classification of daylight and radiation data into three sky conditions by cloud ratio and sunshine duration. Energy Build. 36(7), 660–666. https://doi.org/10.1016/j.enbuild.2004.01.012.

Salcedo-Sanz, S., Cornejo-Bueno, L., Prieto, L., Paredes, D., García-Herrera, R., 2018. Feature selection in machine learning prediction systems for renewable energy applications. Renewable Sustainable Energy Rev. 90, 728–741. https://doi.org/10.1016/j.rser.2018.04.008.

Sharp, T.A., Thomas, S.L., Cubuk, E.D., Schoenholz, S.S., Srolovitz, D.J., Liu, A.J., 2018. Machine learning determination of atomic dynamics at grain boundaries. Proc. Natl. Acad. Sci. 115 (43), 10943–10947.

Solorio-Fernández, S., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., 2019. A review of unsupervised feature selection methods. Artif. Intell. Rev. 53 (2), 907–948. https://doi.org/10.1007/s10462-019-09682-y.

Suárez-García, A., Granados-López, D., González-Peña, D., Díez-Mediavilla, M., Alonso-Tristán, C., 2018. Seasonal caracterization of CIE standard sky types above Burgos, northwestern Spain. Sol. Energy 169, 24–33. https://doi.org/10.1016/j.solener.2018.04.028.

Torres, J.L., de Blas, M., García, A., Gracia, A., de Francisco, A., 2010a. Sky luminance distribution in Pamplona (Spain) during the summer period. J. Atmos. Sol.-Terres. Phys. 72 (5–6), 382–388. https://doi.org/10.1016/j.jastp.2009.12.005.

Torres, J.L., de Blas, M., García, A., Gracia, A., de Francisco, A., 2010b. Sky luminance distribution in the North of Iberian Peninsula during winter. J. Atmos. Sol.-Terr. Phys. 72 (16), 1147–1154. https://doi.org/10.1016/j.jastp.2010.07.001.

Tregenza, P.R., 1999. Standard skies for maritime climates. Light. Res. Technol. 31 (3), 97–106. https://doi.org/10.1177/096032719903100304.

Tregenza, P.R., 2004. Analysing sky luminance scans to obtain frequency distributions of CIE Standard General Skies. Lighting Res. Technol. 36 (4), 271–279. https://doi.org/10.1191/1477153504li117oa.

Uetani, Y., Aydinli, S., Joukoff, A., Kendrick, J.D., Kittler, R., Koga, Y., 2003. BS ISO 15469:2004. Spatial distribution of daylight-CIE standard general sky. Vienna, Austria.

Umemiya, N., Kanou, T., 2008. Classification of sky conditions by the ranges of insolation indices considering CIE standard for general sky. J. Light Vis. Environ. 32 (1), 14–19. https://doi.org/10.2150/jlve.32.14.

Visa, S., Ramsay, B., Ralescu, A.L., Van Der Knaap, E., 2011. Confusion matrix-based feature selection. MAICS 710, 120–127.

Wald, R., Khoshgoftaar, T.M., Napolitano, A., 2014. Optimizing wrapper-based feature selection for use on bioinformatics data. The Twenty-Seventh International Flairs Conference.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V., 2001. Feature selection for SVMs. Adv. Neural Inf. Process. Syst. 668–674.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data mining: practical machine learning tools and techniques.

Wong, S.L., Wan, K.K.W., Li, D.H.W., Lam, J.C., 2012. Generation of typical weather years with identified standard skies for Hong Kong. Build. Environ. 56, 321–328. https://doi.org/10.1016/j.buildenv.2012.04.003.

Yang, Y., Pedersen, J.O., 1997. A comparative study on feature selection in text categorization. Icml, Nashville, TN, USA, p. 35.

Yu, L., Liu, H., 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning (ICML-03). pp. 856–863.

Zi, Y., Sun, C., Han, Y., 2020. Sky type classification in Harbin during winter. J. Asian Archit. Build. Eng. 1–12 https://doi.org/10.1080/13467581.2020.1752217.