Computer Methods and Programs in Biomedicine 245 (2024) 108033

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

An explainable machine learning approach for hospital emergency department visits forecasting using continuous training and multi-model regression

C. Peláez-Rodríguez [a,*], R. Torres-López [a], J. Pérez-Aracil [a], N. López-Laguna [b], S. Sánchez-Rodríguez [c], S. Salcedo-Sanz [a]

[a] Department of Signal Processing and Communications, Universidad de Alcalá, Alcalá de Henares, 28805, Spain
[b] Emergency Department, Clínica Universidad de Navarra-Madrid, Madrid, 28027, Spain
[c] Operations Department, Clínica Universidad de Navarra-Madrid, Madrid, 28027, Spain

ARTICLE INFO

MSC:
0000
1111

Keywords:
Hospital emergency departments
Admissions forecast
Machine learning
Multi-step forecasting
Continuous-training algorithms

ABSTRACT

Background and Objective: In the last years, the Emergency Department (ED) has become an important source of admissions for hospitals. Since late 90s, the number of ED visits has been steadily increasing, and since Covid19 pandemic this trend has been much stronger. Accurate prediction of ED visits, even for moderate forecasting time-horizons, can definitively improve operational efficiency, quality of care, and patient outcomes in hospitals.
Methods: In this paper we propose two different interpretable approaches, based on Machine Learning algorithms, to accurately forecast hospital emergency visits. The proposed approaches involve a first step of data segmentation based on two different criteria, depending on the approach considered: first, a threshold-based strategy is adopted, where data is divided depending on the value of specific predictor variables. In a second approach, a cluster-based ensemble learning is proposed, in such a way that a clustering algorithm is applied to the training dataset, and ML models are then trained for each cluster.
Results: The two proposed methodologies have been evaluated in real data from two hospital ED visits datasets in Spain. We have shown that the proposed approaches are able to obtain accurate ED visits forecasting, in short-term and also long-term prediction time-horizons up to one week, improving the efficiency of alternative prediction methods for this problem.
Conclusions: The proposed forecasting approaches have a strong emphasis on providing explainability to the problem. An analysis on which variables govern the problem and are pivotal for obtaining accurate predictions is finally carried out and included in the discussion of the paper.

1. Introduction

The forecasting of Emergency Department (ED) visits is extremely important for improving hospital's operations, in terms of healthcare management and resources allocation, given the prevalent global challenge of ED overcrowding [1–5]. Accurate predictions of patient demand in the ED can greatly enhance the operational efficiency, quality of care, and patient outcomes in hospitals [6–9]. By anticipating the volume and characteristics of ED visits, healthcare administrators and clinicians can proactively allocate resources such as staff, equipment, and treatment facilities to match the anticipated demand. Moreover, these data have been even used in the past to predict epidemics [10]. It has been also proven that an efficient forecasting of ED visits leads to more efficient scheduling, timely provision of care, and effective response to surges in patient influx [11,12]. Moreover, accurately forecasting ED visits has an impact in optimizing resource utilization, reducing waiting times, and improving patient satisfaction [13,14].

Extensive research has been conducted in last years on forecasting the patient volume expected to visit a hospital's ED during operational hours [15–17]. Within the most frequently employed prediction methods, the utilization of parametric methods based on linear regression [18–20] for the analysis of time series, stands out as the prominent approaches. Time series forecasting methods refer to analytical techniques employed to predict future values or patterns in a time-ordered

* Corresponding author.
E-mail address: cesar.pelaez@uah.es (C. Peláez-Rodríguez).

https://doi.org/10.1016/j.cmpb.2024.108033
Received 30 October 2023; Received in revised form 8 January 2024; Accepted 14 January 2024
Available online 23 January 2024

sequence of observations based on past data [21]. Time series models used for forecasting include autoregressive moving average (ARIMA) [22–25], exponential smoothing [26,27] or seasonal autoregressive integrated moving average (SARIMA) [28–30,26,18]. One of the main drawbacks of these models resides in the fact that time series models exclusively rely on the available data of the variable being forecasted, without explicitly considering the underlying factors influencing its behavior. These models primarily extrapolate trends and seasonal patterns while disregarding additional information, such as weather conditions or calendrical variables. To prevent this issue, multivariate methods have also been adopted, such as multivariate autoregressive integrated moving average (MSARIMA) [29,30], seasonal autoregressive integrated moving average with external regressor (SARIMAX) [31] or vector autoregression (VAR) [32]. Also, other approaches based on discrete event simulations of ED patient flow for forecasting near-future operation conditions have been proposed to this end [33,34]

Due to the stochastic and non-linear nature of ED flow, there has been a growing trend towards the utilization of non-parametric methods in the past decade. Thus, several Machine Learning (ML) based forecasting algorithms have been implemented for the prediction of ED visits, and also in telemedicine emergencies [35], employing different types of predictive variables and various levels of time granularity. In [36] ML models are presented, which tackle the real-time and personalized waiting time prediction in an ED. In [37] the daily patient arrivals at an Emergency Department (ED) were modeled using an Artificial Neural Network (ANN), with the objective of quantifying the importance of different contributing variables. Also, in [38] data regarding weather, days of week, air quality, and special events were used to train the ANN for predicting ED volume. In [39] a multilayer perceptron (MLP) model was used to forecast future daily arrival and hourly occupancy levels, given recent history and useful exogenous variables. Other ML techniques as Decision Trees(DT) [40], Random Forests (RF) [41–43], Gradient Boosted Machines (GBM) [42], Support Vector Regression (SVR) [44,45] or XGBoost [46,47,43,48] have also been employed for the prediction of ED visits.

Within the ML forecasting algorithms, Deep Learning (DL) approaches have gained growing significance in the last decade. These methods are composed of multiple processing layers to learn with multiple levels of abstraction, and have become a very popular alternative to traditional methods in the field of time series forecasting. In [49] a total of eight different DL models have been compared for one- and multi-step-ahead ED visits forecasting using two real datasets from CHRU Lille, France. In [11] a deep stacked architecture is being proposed and applied to the daily ED visits prediction problem with deep components such as Long Short Term Memory (LSTM), Gated Recurrent Units (GRU) and simple Recurrent Neural Network (RNN). In [49] a variational autoencoder (VAE) model to forecast daily and hourly visits at a ED with two-year data was developed. DL approaches based on RNN [50], LSTM [41,51] or Convolutional Neural Network (CNN) [41] can be also found in the literature.

One of the intrinsic difficulties of the ED visits forecasting problem resides in the unpredictable nature of medical emergencies. Patient flow in a hospital can be highly dynamic and subject to constant changes, due to various factors such as seasons, special events, disease outbreaks, or socio-economic conditions, among others. Therefore, since time series forecasting relies on capturing patterns and relationships in historical data to make predictions about future behavior, the fact that the training data represent obsolete trends that do not correspond to the dynamics of the current time series may pose a problem. This fact may lead to the production of inaccurate or unreliable predictions. To address this issue, regularly updating the training data with the most recent observations helps capture the current trends and dynamics of the time series, improving the performance of the forecasts. Taking this into account, a forecasting approach based on continuous training has been implemented in this work, aiming at improving the accuracy of the predictions. Continuous training (also known as online learning or incremental learning in the literature) is a ML approach that involves updating and refining a model over time, as new data becomes available [52,53]. In continuous training, the model is first trained on an initial dataset, and then it is incrementally updated as new data is collected. This ongoing training process enables the model to capture changes and trends in the data, adapt to evolving patterns, and stay up-to-date with changing conditions, leading to more accurate and reliable predictions and insights. Overall, continuous training offers a powerful framework for building adaptive and responsive ML models that can continuously learn from new data and improve their performance over time. Although this technique has been applied to some time series related forecasting problems such as wind power prediction [54], smart agriculture [55] or smart traffic management [56], this is the first time, to the best of the authors' knowledge, that its relevance is assessed in the prediction of hospital ED visits.

While the previously mentioned regression methods can effectively forecast the number of patient visits with a reasonable precision, it is equally important to explain the individual decisions of policy makers. Explainability fosters trust and acceptance of the forecasting model among healthcare professionals, stakeholders, and the general public. When the underlying mechanisms and reasoning of the forecasts are transparent, it becomes easier for individuals to understand and trust the predictions. This can facilitate adopting and implementing the forecasting model in real-world settings. In this context, the great impact of using explainable AI (XAI) techniques has become increasingly evident. XAI refers to the ability of AI systems to provide transparent explanations of their predictions and decision-making processes [57]. Although there is no extensive literature on explainable forecasting methods in ED visit prediction problems, existing research has shown a strong association between weather-related variables, such as daily temperature and precipitation, and the volume of patient visits in healthcare facilities [58,40]. Also, in [46,59] Shapley Additive exPlanations (SHAP) is applied to interpret the impact of variables on the number of patient visits. Also, explainability has been tackled in other problems related human-based activities such as bike sharing [60] or school absences [61].

Seeking for an improvement in the performance of the predictive models, while ensuring transparency in the prediction process, a data segmentation strategy has been adopted in this work. This strategy involves partitioning the database into multiple subsets based on different criteria, and distinct ML models are then trained separately on each subset. This approach aims to achieve improved results by training models tailored to each data subset's unique characteristics. Simultaneously, it enables the analysis of how predictor variable distributions vary across different groups, and how this variation influences the prediction model performance. This field of the ML area is known as *ensemble regression*, and has gained a popularity in recent years [62]. Similar architectures have been proposed for solving different regression problems, e.g.: wind power is forecasted using cluster-based ensemble learning in [63,64]. There are also some examples that can be found in the literature where these techniques are applied to the problem of ED visits forecasting. In [65] a dataset of four years of ED attendances is split into seven according to the weekday, and ARIMA, neural networks and Fuzzy Time Series (FTS) forecasting models are applied to each subset independently. In [66] a new methodology to determine the relative importance of the predictor variables (XAIRE) is presented, it aggregates the relative importance of each predictor variable obtained by each regression method, obtaining a general ranking.

In this paper we introduce, compare, and analyze two different ML-based interpretable approaches, that rely on continuous training and ensemble regression, to address the challenge of forecasting hospital ED visits. The proposed approaches involve data segmentation based on two criteria: In the first one, a threshold-based strategy is adopted, where data are divided depending on the value of specific predictor variables. This allows us to assess which variables and threshold combinations yield superior forecasting results. In the second approach, a

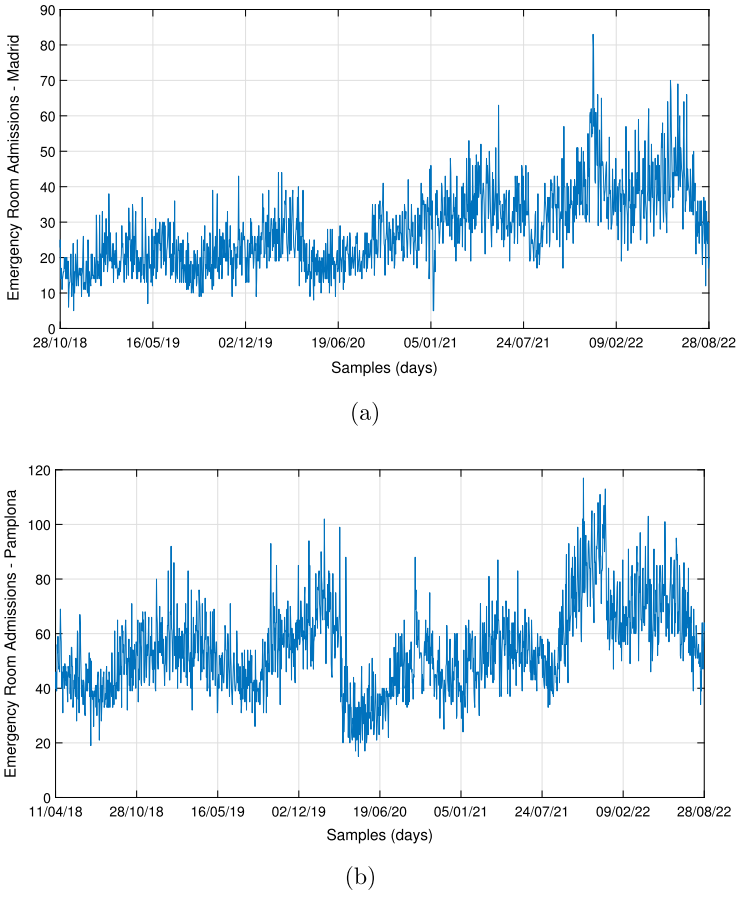

Fig. 1. Emergency rooms admissions time series for Madrid and Pamplona hospitals considered; (a) Madrid; (b) Pamplona.

cluster-based ML ensemble learning is proposed. In this case, a clustering algorithm is applied to the training dataset, and ML models are then trained for each cluster. Finally, when predicting an incoming sample, we determine the nearest cluster for that sample, and utilize the corresponding ML model to generate the prediction. These methodologies are evaluated and compared using real ED visits datasets from two hospitals in Spain, demonstrating their remarkable efficacy compared to different baseline ML models.

The rest of the manuscript has been organized as follows: Section 2 describes the admission to the ED datasets used in this paper, and the proposed methodologies implemented throughout this work. Section 3 presents the experimental work carried out, and reports and analyzes the results. Finally, Section 4 closes the paper with some final remarks and conclusions.

2. Methods

2.1. Data description and analysis

In this section, we present and analyze the two datasets employed for the ED visits predictions carried out. We have obtained data of admissions to the ED from Clínica Universidad de Navarra, a private institution in Spain with two hospitals, one in Pamplona (Navarre) where it has its headquarters, and another one in Madrid city. Clínica de Universidad de Navarra is a high-resolution hospital characterized by its diagnostic speed, owing to its multidisciplinary work and the acquisition of the latest technology to offer care in 46 different medical and surgical specialties in the first 24–48 hours, and treatment within the first 72 hours. To meet those speed and quality of care requirements, the clinic must optimize all the processes involved from the moment the patient comes to the hospital, especially when it happens in the emergency room, where appointments do not exist, and waiting time could be critical for the treatments.

The ED is usually divided into three emergency rooms and protocols: General admissions, pediatrics and gynecology. According to the doctor in charge of the ED in Madrid, the admissions in pediatrics follow a different pattern from the admissions in the general room, which are generally more correlated with the day of the week. Additionally, the number of admissions in gynecology was not relevant, since many of births are previously planned nowadays, and they follow a different protocol in the ED with specific specialists. For this reason, in our research, we have only considered General admissions. The hospital in Madrid started its emergency service in September 2018 and since then, the number of admissions has had a remarkable increase. To draw consistent conclusions for comparison, we have also considered the number of admissions in Pamplona since January 2018, although its emergency service started much earlier. Note that including data from two different hospitals, one in Pamplona (North of Spain) and another one in Madrid (center of Spain), allows us to discuss ED visits prediction in two similar hospitals (same institution), but located in very different cities in Spain, so different patterns in ED visits prediction can appear, and the proposed algorithms should be able to adapt to each case.

In both cases, a prediction problem can be mathematically formulated as a regression problem, in which the number of admissions to the emergency room must be estimated. The original data contains a time tag in which the admission occurs, but for simplicity, we have aggregated this information to perform a daily time series. The two datasets can be accessed through the following public repository: https://github.com/RTLPHD/EmergencyDepartmentForecasting.

Fig. 1 shows the time series of the two problems considered, with details on the dates ranges in which the data are available. In both cases, the effect of Covid-19 can be seen from March 2020 to September 2020,

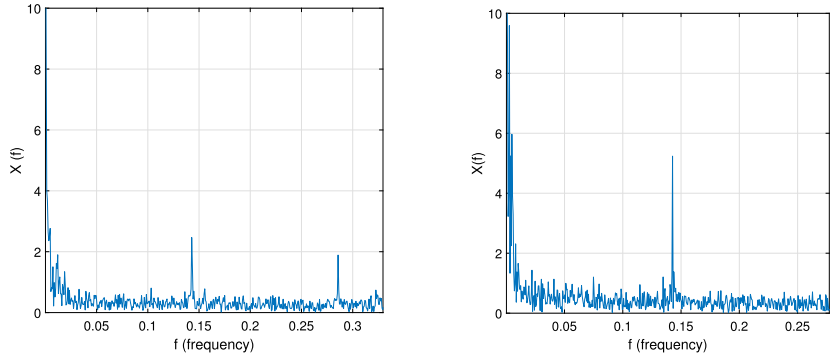

Fig. 2. Emergency room admissions (FFT series) - Madrid and Pamplona hospitals; (a) Madrid; (b) Pamplona.

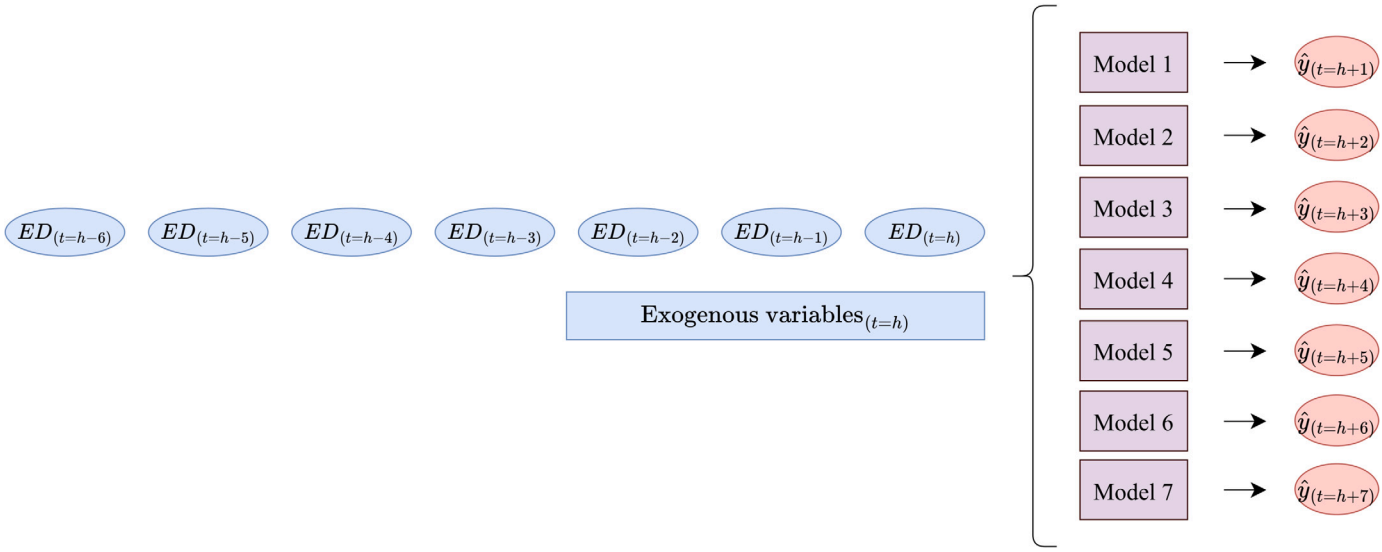

Fig. 3. Direct multi-step forecast strategy employed to predict the volume of ED visits over the entire week.

when the number of admissions suddenly decreased: admissions due to Covid were treated with specific protocols, and those figures were not included in the datasets.

When we calculate and represent the Fast-Fourier-Transform (FFT) of the admissions time series (Fig. 2) to locate admission frequencies in the series, it is possible to see a remarkable peak at 7-days frequency (0.143) in both time series of emergency admissions, showing a clear weekly frequency of the admissions in these hospitals. In Madrid there is also a clear peak in 3.5 days (half week) $f = 0.285$ which does not seem to clearly appear in Pamplona. There are also other peaks at the lowest frequencies in both cases, easier to identify in Pamplona (Navarre) figure, that corresponds to annual frequency (0.00269).

2.2. Proposed methodology

This section details the proposed methodology to predict emergency department visits. Initially, the ML models used throughout the paper are summarized in Section 2.2.1. Then, the multi-step forecasting strategy for obtaining the weekly ED visits forecast is detailed in Section 2.2.2. Subsequently, the continuous training approach used to improve the performance of the base models in the weekly ED visits forecast is further described in Section 2.2.3. Finally, Section 2.2.4 provides details about the data segmentation approach that allows the extraction of explainable conclusions for the forecasting problem at hand.

2.2.1. Regression models

Six different regression methods have been selected for carrying out the ED visits forecasting. They include classical shallow ML methods such as Linear Regression [67], Regression Trees [68], Random Forest [69] and Support Vector Regression [70], and two different architectures of artificial neural networks: Extreme Learning Machine [71] and Fully Connected Deep Neural Network [72].

2.2.2. Multi-step forecast strategy

Within this framework, a direct multi-step forecast strategy has been employed to predict the volume of ED visits over a week. This methodology needs the creation of individual models for each forecasting time step. As a result, each model generates a different output while utilizing the same input predictors. This approach is depicted in Fig. 3.

2.2.3. Continuous training

Then, a continuous training approach has been implemented with the aim of improving the multi-step prediction provided by the regression model. Also, this approach has the objective of overcoming the intrinsic difficulty of this type of highly dynamic regression problems, where it is easy that past training data represent obsolete trends that do not correspond to the dynamics of the current time series. Continuous training, constitutes a ML technique that entails iteratively updating and refining a model over time, as new data becomes accessible.

The schematic representation of the proposed approach is depicted in Fig. 4. The procedure starts with the construction of the initial

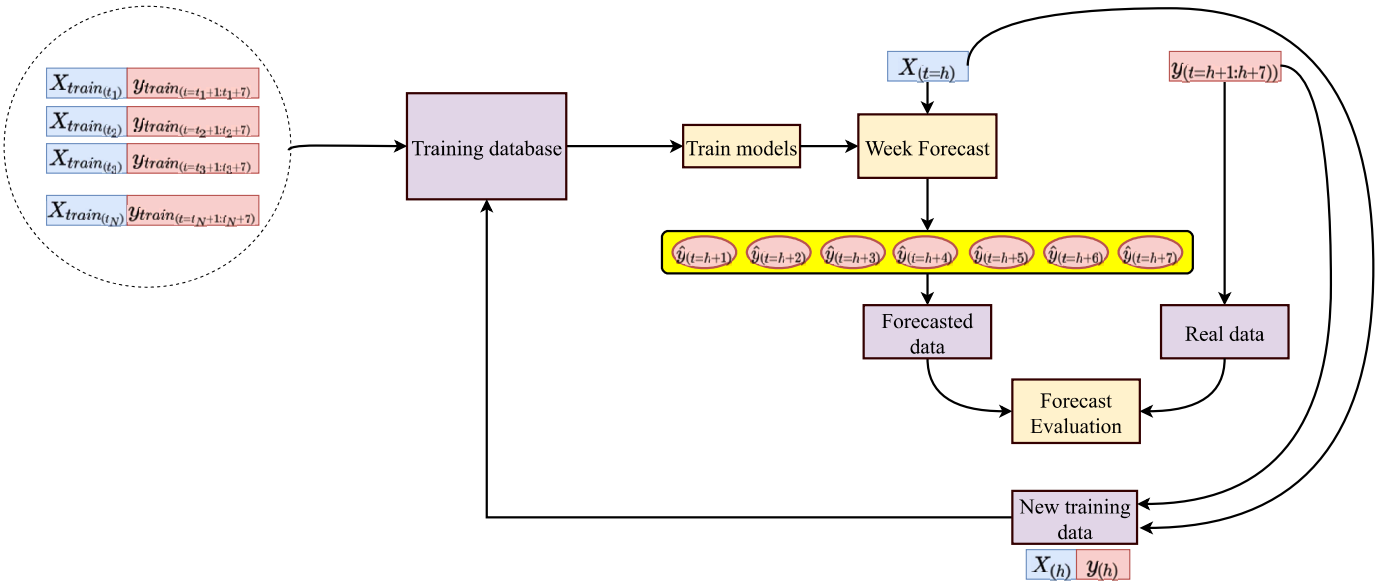

Fig. 4. Continuous training proposed model.

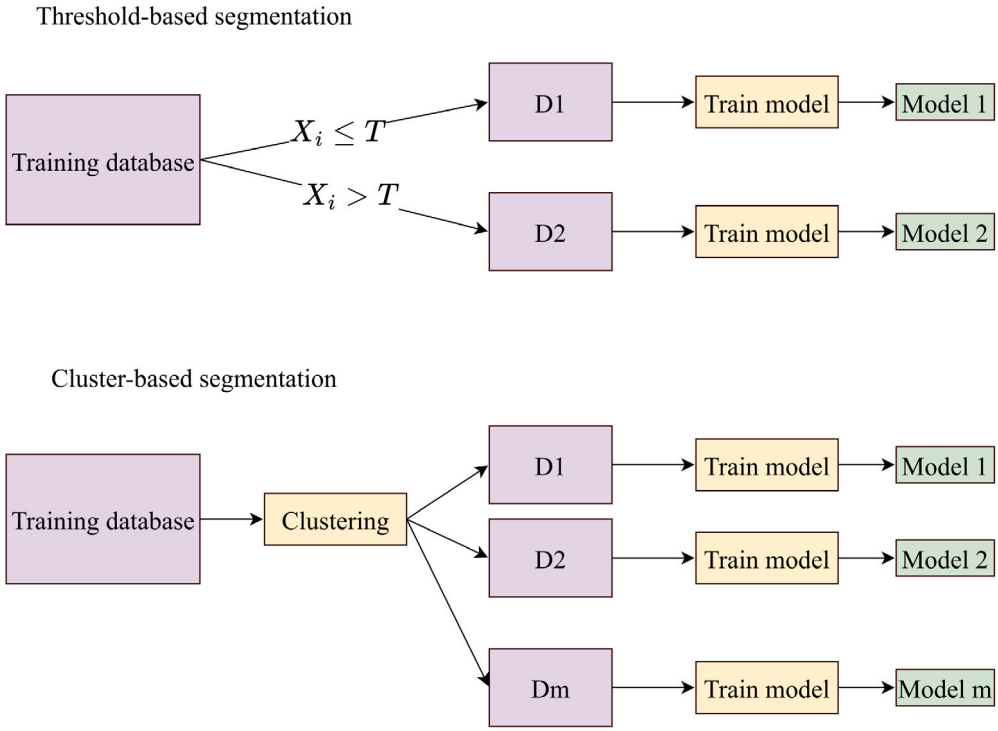

Fig. 5. Flowchart depicting the two data segmentation strategies considered.

training database utilizing the accessible training data. Subsequently, regression models are trained using this compiled dataset. Then, as a novel test sample is introduced, the ML models predict the anticipated volume of ED visits for the ensuing week. The accuracy of these predictions is later evaluated against actual observed data. In the final step, the factual data encompassing both predictors and target variables are assimilated into the training database, effectively expanding it with a new data point. The regression models are then retrained to forecast forthcoming test samples based on the updated information. This cyclic process iterates for each weekly forecast, ensuring the continual refinement and adaptation of the model in response to new data.

This continuous training paradigm presents a robust framework for constructing adaptive ML models capable of consistently assimilating new trends of data and improve their performance over time, as it is demonstrated in the results section.

2.2.4. Data segmentation

In addition to the aforementioned approach, and with the objective of enhancing predictive performance, while offering interpretability to the prediction problem, a data segmentation methodology has been included in the system. In this approach, the training database is divided based on a specific criterion, resulting in the training of two or more models for predicting each target variable. Fig. 5 illustrates the operation scheme of the training database segmentation process for the subsequent training of the individual models. The specific criteria employed for data segmentation are detailed in Sections 2.2.4.1 and

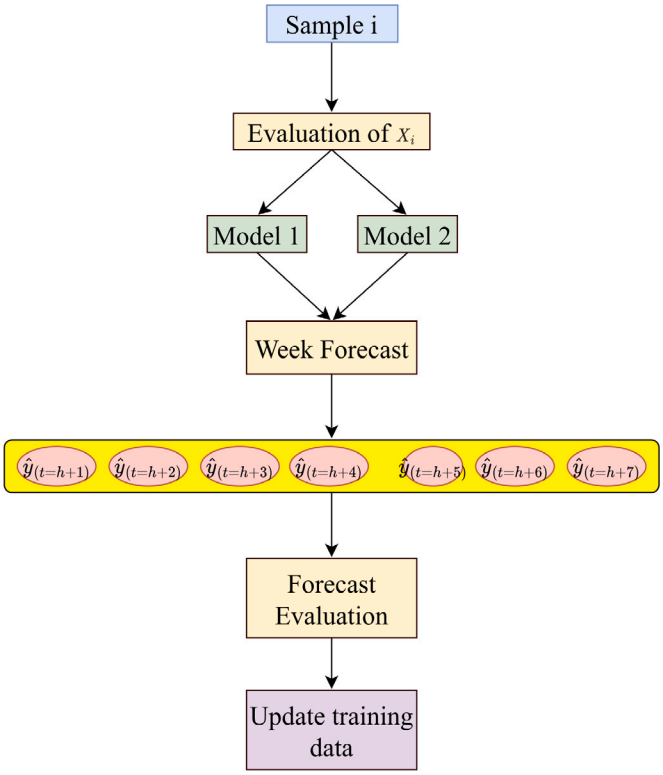

Fig. 6. Prediction procedure after the training data segmentation has been performed.

2.2.4.2. Then, Fig. 6 depicts the prediction process of a new sample, where the assignment of a model is performed upon the evaluation of the predictor variables values. Subsequently, the forecast value is derived from the designated model.

2.2.4.1. Threshold-based segmentation Threshold-based segmentation involves the division of a dataset into different subsets or segments, based on specific threshold values set for one predictor variable (X_i). The primary objective of this segmentation approach is to group data points that exhibit similar characteristics or behavior, according to the chosen threshold criteria. By doing so, it enables the creation of multiple models, each tailored to a particular subset of the data. This strategy facilitates improved predictive accuracy and explainability of the prediction model. Indeed, the versatility of threshold-based segmentation allows for an in-depth exploration of prediction performance by experimenting with various predictor variables and threshold values. This approach provides the opportunity to assess the significance of different variables and thresholds in relation to the specific problem under investigation. This comprehensive analysis helps refine predictive models, uncover hidden patterns, and ultimately enhances the understanding and decision-making capabilities in the context of the studied problem.

An ensemble approach has been adopted to compute the ultimate prediction for an incoming sample. This methodology involves calculating the predicted value by individually utilizing each predictor variable for segmentation. Specifically, percentiles at several levels (P5, P10, P15, …, P85, P90, P95) of each variable have been employed as threshold values in the segmentation process. Then, the final predicted value is computed as the average value of all the combinations of variables and thresholds considered. By employing this ensemble strategy, a comprehensive exploration of different variable-threshold combinations has been conducted, to ensure robust and well-informed predictions, enhancing the predictive accuracy and adaptability of the model.

2.2.4.2. Cluster-based segmentation Cluster-based segmentation is a data partitioning technique that is based on the concept of grouping data points into clusters, where each cluster represents a subset of data with shared characteristics or similarities. The K-means algorithm [73] is designed to perform the data segmentation. It is designed to partition data (N samples each having measurements on P variables) into K classes $(C_1, C_2, \cdots, C_K)$, where C_k is the set of n_k objects in cluster k, and K is given. It operates by trying to separate samples in K groups of equal variance, each characterized by a centroid representing the mean of the samples within that cluster (μ_j), minimizing a criterion known as the inertia or within-cluster sum-of-squares:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\left\| x_i - \mu_j \right\|^2 \right) \tag{1}$$

By minimizing inertia, K-means seeks to find the optimal arrangement of clusters, effectively separating the samples into coherent and internally homogeneous groups. Then, in the prediction phase, incoming samples are assigned to their respective clusters based on the proximity to cluster centroids. Each sample's distance to all centroids is calculated, and the sample is assigned to the cluster with the closest centroid.

As in the previous case, an ensemble-based approach is employed to maximize the information extraction. In this approach, a range of cluster numbers, spanning from 2 to 10, is defined. For each cluster configuration, the ED prediction is calculated, and the final prediction is derived as the average of these individual predictions. This ensemble strategy harnesses the predictive power of various cluster configurations, yielding a more robust and informed final prediction, thus enhancing the accuracy and reliability of the model's forecasts.

3. Experiments and results

This section presents and discusses the outcomes obtained during the course of this study. First, Section 3.1 details the data pre-processing steps performed. After that, Section 3.3 shows the base results obtained with the ML models considered. Then, Section 3.4 presents the improved results after applying the continuous training approach. Subsequently, results and conclusions extracted after performing the data segmentation techniques are reported in Section 3.5. Finally, a comparison among methods and strategies is conducted in Section 3.6.

3.1. Experimental set-up

The initial phase of the experimental work involved preparing the database through the following pre-processing steps: (1) database construction (Section 3.1.1), (2) features scaling (Section 3.1.2), (3) training-test split (Section 3.1.3). Then, details of the specific regression metric used for assessed the performance of the models are provided in Section 3.2. Finally, Section 3.2.1 describes the hyperparameters used regression models tuning.

3.1.1. Database construction

The process of constructing the database has involved the incorporation of three categories of predictive variables:

- Meteorological variables: information regarding temperature, pressure and rainfall for the three previous days.
- Categorical/Calendrical context variables: variables related to the day of week, moon phase, and the presence of festivities on both the current and preceding days.
- Persistence related variables: information related to the autoregressive pattern, corresponding to the daily ED visits of the previous week.

Regarding the target variable, for each sample of the database seven independent target values are considered, corresponding to the daily ED visits values for the next week, as indicated in Fig. 3.

The specific predictive variables considered for each problem are listed in Figs. 7 and 8 for Pamplona and Madrid cases, respectively. In

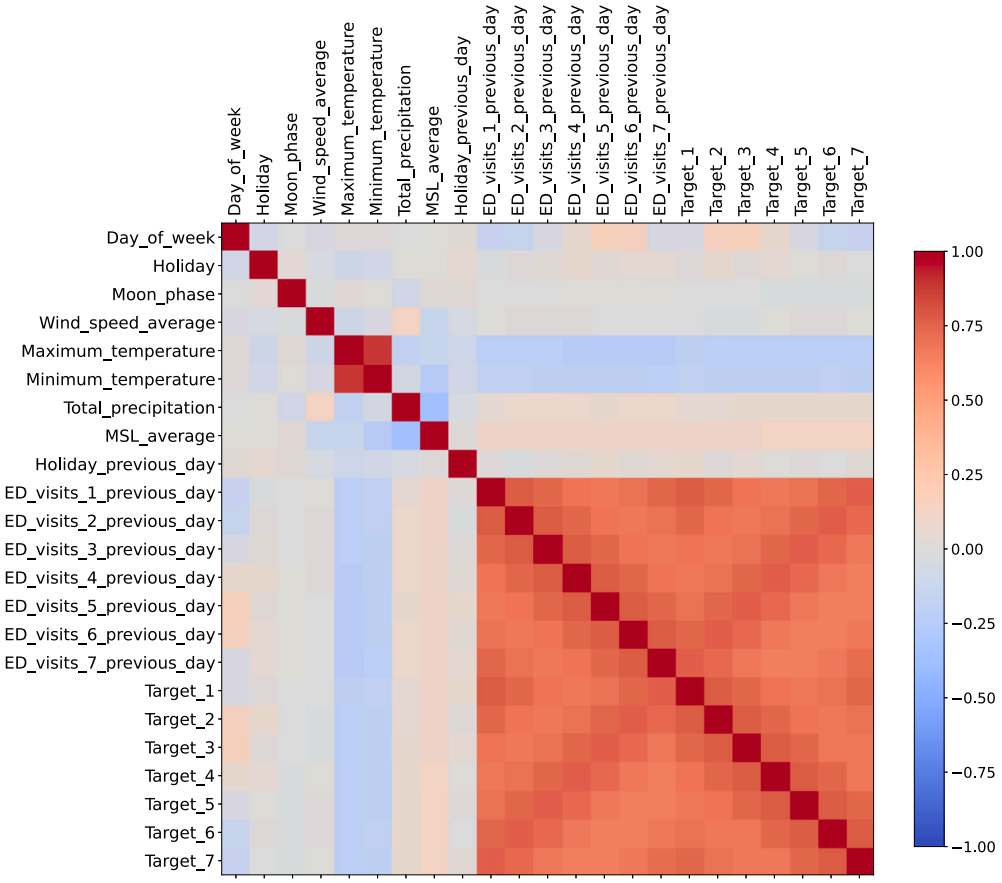

Fig. 7. Correlation coefficients among the variables belonging to Pamplona dataset.

these Figures, the correlation between variables is depicted, both among predictor variables and target variables. In the case of meteorological variables, the average, maximum or minimum labels indicate the corresponding value computed in the three previous day of the prediction, i.e.: if we want to forecast the daily ED visits from day h to day $h+6$, meteorological values are computed from day $h-3$ to day $h-1$.

When examining the correlation heat-maps, similar observations can be made for both of the investigated scenarios. First, variables related to persistence demonstrate the strongest correlation levels with respect to the target variables. Second, concerning meteorological variables, it becomes evident that temperature-related variables exhibit a negative correlation with ED visit occurrences, indicating that higher temperatures are associated with fewer ED visits in both hospitals. Additionally, precipitation and pressure variables display a slight positive correlation, while wind speed variables exhibit no correlation in either of the databases. Finally, limited conclusions can be drawn regarding the correlation between calendrical and target variables, as this set of predictors exhibits correlation values close to zero in both cases.

3.1.2. Feature scaling

Scaling the features represent an important step when approaching a prediction problem, in order to ensure that the upper and lower limits of data are in a given predefined range. Feature standardization was performed in this paper, causing data to have zero-mean and a unit-variance (Equation (2)), as follows:

$$x' = \frac{x - \bar{x}}{\sigma} \tag{2}$$

where x is the original feature vector, $\bar{x}$ denotes the feature mean and σ its standard deviation.

3.1.3. Training-test split

Finally, a training-test split (70%-30%), considering the in-sample and out-of-sample was considered, assuring that no test instance was seen by the ML method during the training process. Since we deal with timed-series data (instead of randomly splitting the datasets), the last 30% of instances have been removed to validate the methods. Figs. 9 and 10 display the time series of ED visits for the two studied cases after the training-test split was performed. It is noticeable in these figures that a challenge addressed in this study arises, as the test dataset exhibits considerably higher values of ED visits compared to those observed in the training dataset.

3.2. Regression metric

A commonly used regression metric has been employed to assess the performance of the ML methods applied to the two proposed problems: the Mean Absolute Error (MAE), whose corresponding equation is:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

where $\hat{y}$ represents predicted values (provided by the model) and y are the actual values. The subscript i is used to refer to a single sample $y_i = y[i]$.

3.2.1. Hyperparameters tunning

Given the significant importance of selecting optimal hyperparameters to ensure the effective operation of ML models, a randomized search hyperparameter tuning was performed, where a K-fold cross validation with $K = 5$ was selected and a number of 50 iterations were set. The specific hyperparameters considered for each model and its corresponding search ranges are listed in Table 1.

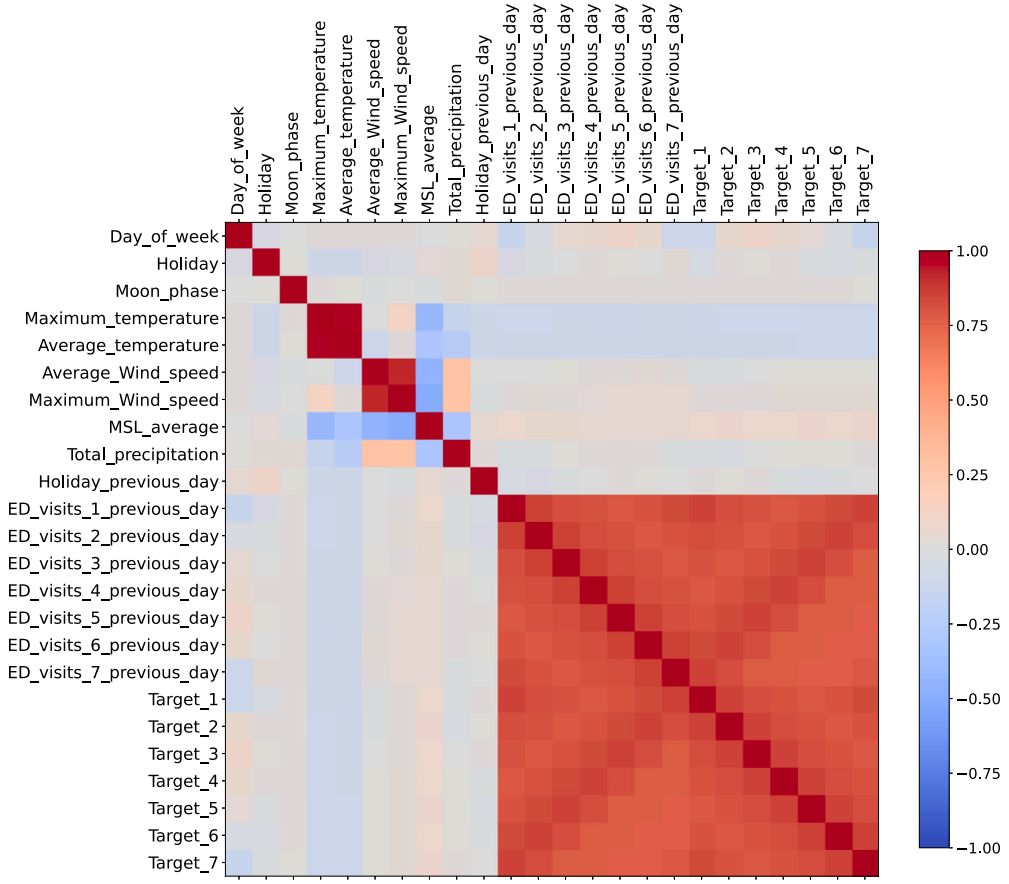

Fig. 8. Correlation coefficients among the variables belonging to Madrid dataset.

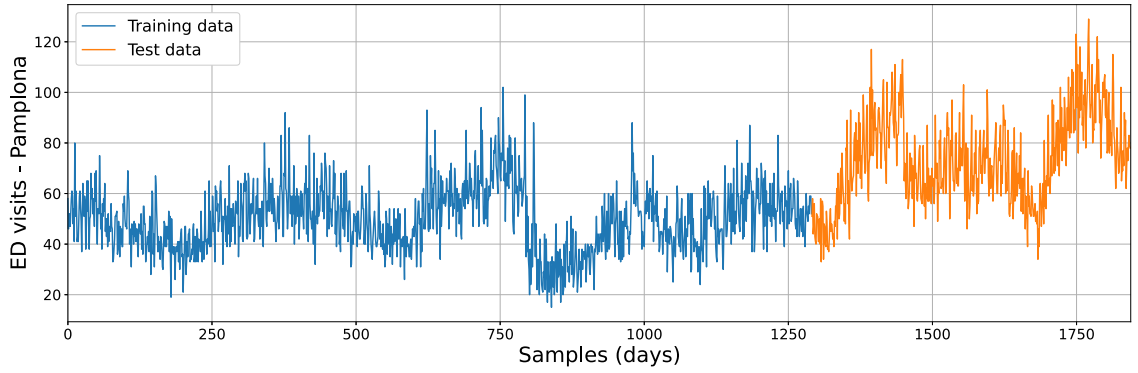

Fig. 9. Time series of Pamplona ED visits.

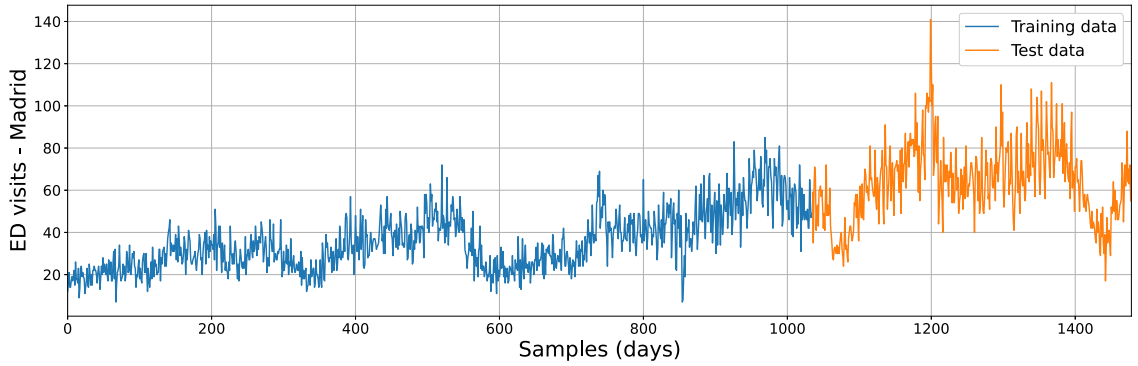

Fig. 10. Time series of Madrid ED visits.

Table 1
Experimental setup.

RT		RF	
Max. depth	1-20	Max. depth	1-20
Min. leaf	1-20	Min. leaf	1-20
Criterion	mae, mse, friedman mse	Criterion	mae, mse, friedman mse
min samples split	2	Nº of estimators	1-500
min weight fraction leaf	0	Max. features	2-21
		Bootstrap	True, False
SVR		**ELM**	
C	1-100	Nº of hidden neurons	100-1000
Kernel	linear, rbf		
Gamma	10^{-4}-10^{-2}		
Epsilon	0-1		
Tolerance	10^{-4}-10^{-1}		
FCDNN			
Nº of layers	1-3		
Neurons per layer	10-200		
Activation	relu, sigmoid, tanh		
Epochs	100,200,300,400,500		
Batch size	16,32,64		

Table 2
Base results (MAE) of Pamplona ED visits forecasting.

	+1 day	+2 days	+3 days	+4 days	+5 days	+6 days	+7 days
LR	9.00	9.29	9.81	10.20	10.49	10.85	10.89
RT	10.74	16.89	13.85	12.46	16.43	17.01	14.21
RF	12.64	12.34	13.79	13.60	13.70	15.94	14.83
SVR	9.09	9.34	9.63	9.97	10.50	10.75	10.84
ELM	8.83	9.34	9.84	10.40	10.89	10.77	11.20
FCDNN	**8.70**	**8.79**	**8.79**	**9.14**	**9.29**	**9.58**	**10.12**
Average	9.83	11.00	10.95	10.96	11.88	12.48	12.02

Table 3
Base results (MAE) of Madrid ED visits forecasting.

	+1 day	+2 days	+3 days	+4 days	+5 days	+6 days	+7 days
LR	**8.54**	9.04	9.26	**9.54**	**9.65**	9.77	9.87
RT	10.60	11.58	11.55	11.49	11.69	11.27	12.03
RF	10.60	10.81	11.07	10.79	10.73	11.21	11.41
SVR	8.59	9.02	**9.25**	9.59	**9.65**	**9.66**	9.93
ELM	8.67	**8.91**	9.41	9.82	9.77	9.80	**9.83**
FCDNN	8.67	8.93	9.29	9.71	9.98	9.70	9.94
Average	9.28	9.72	9.97	10.16	10.25	10.24	10.50

3.3. Base results

After tuning and training the six ML methods with the training data, their predictive performance is assessed using the test data. Tables 2 and 3 present the outcomes of these models for each of the databases under consideration, encompassing the results for all the target variables in terms of MAE. It is important to bear in mind that each target variable is predicted using a separate regression model.

Two conclusions can be drawn from these initial results: (1) As expected, the predictive performance deteriorates as the prediction time horizon increases, leading to a decline in the average results by up to 26.96% for Pamplona database and 13.15% for Madrid database; and (2) regarding the performance of specific ML models, it is observed that four of the six models (LR, SVR, ELM, and FCDNN) produce remarkably similar results, while the CART models (RT and RF) exhibit notably inferior performance.

3.4. Continuous training

Next, in order to improve the predictive performance of the models, making them able to adapt to emerging data trends observed in the test data, the continuous training scheme described in Section 2.2.3 has been implemented. This approach involves, once the prediction of an incoming test sample is made, incorporating it into the training data and retraining the model, allowing the model to learn from the new data.

The results obtained after this approach are performed are reported in Tables 4 and 5 for Pamplona and Madrid databases, respectively. In these tables, the significant improvement of the average results of the six ML models is observed regarding the base results shown in Tables 2 and 3, reducing the error metrics between 8.34% and 19.63% in Pamplona database, and between 3.74% and 4.88% in Madrid database.

Then, Figs. 11 and 12 provide a comparison of the predictions made by the base and the continuous training models. Figures represent the +1 day predictions for Pamplona and Madrid hospitals, respectively, where X-axis represents actual ED visits and Y-axis denotes the predicted values. For an ideal model, the points should form a line of slope one (shown in black). In both Figures, top row represents the base model (base) and last row represents the continuous training models (retrain). It is possible to observe how the retrain models tend to exhibit notably improved performance, especially on days with a high

Table 4
Results (MAE) for Pamplona database after applying the continuous training approach.

	+1 day	+2 days	+3 days	+4 days	+5 days	+6 days	+7 days
LR	8.76	8.82	9.17	9.34	9.51	9.71	9.62
RT	10.04	11.11	10.64	10.58	11.11	11.29	10.84
RF	8.89	8.97	9.16	**9.12**	**9.34**	9.91	9.80
SVR	8.77	8.88	9.17	9.35	9.55	9.75	9.61
ELM	**8.68**	8.77	9.12	9.52	9.55	9.84	9.74
FCDNN	8.92	**8.66**	**8.94**	9.16	9.56	**9.65**	**9.55**
Average	9.01	9.20	9.37	9.51	9.77	10.03	9.86
Improvement	8.34%	16.36%	14.43%	13.23%	17.76%	19.63%	17.97%

Table 5
Results (MAE) for Madrid database after applying the continuous training approach.

	+1 day	+2 days	+3 days	+4 days	+5 days	+6 days	+7 days
LR	**8.45**	9.01	**9.20**	9.46	**9.51**	9.58	**9.71**
RT	10.25	10.42	10.44	10.77	10.83	10.72	10.88
RF	8.72	9.13	9.59	9.70	9.54	9.76	10.18
SVR	8.53	9.02	9.22	**9.45**	**9.51**	**9.56**	9.79
ELM	8.77	**8.90**	9.25	9.56	9.57	9.63	9.89
FCDNN	8.64	9.01	9.31	9.74	**9.51**	9.68	9.95
Average	8.89	9.25	9.50	9.78	9.75	9.82	10.07
Improvement	4.20%	4.84%	4.71%	3.74%	4.88%	4.10%	4.10%

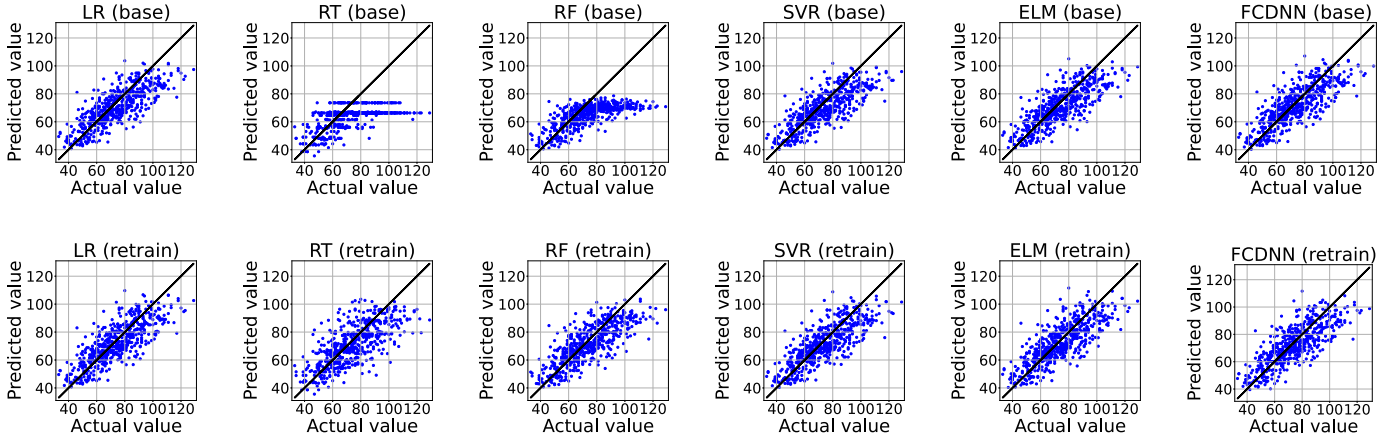

Fig. 11. Scatter plot of the actual vs predicted values for the six ML methods tested for Pamplona database and the +1 day horizon. First row represents the base model while second row represents the continuous training models.

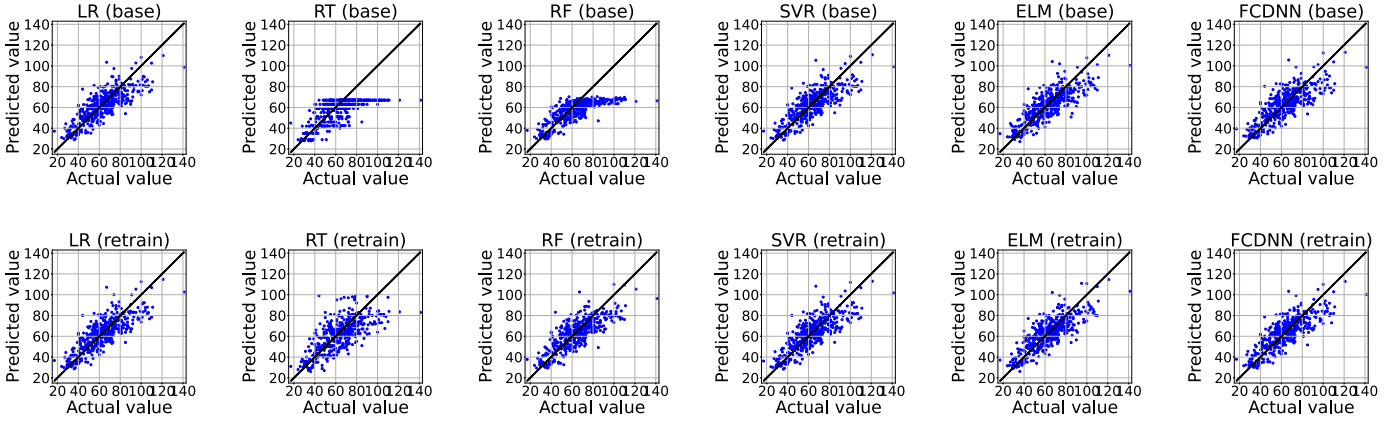

Fig. 12. Scatter plot of the actual vs predicted values for the six ML methods tested for Madrid database and the +1 day horizon. First row represents the base model while second row represents the continuous training models.

Table 6
Results (MAE) for Pamplona database applying the threshold-based segmentation approach.

	+1 day	+2 days	+3 days	+4 days	+5 days	+6 days	+7 days
LR	8.72	8.77	9.15	9.31	9.48	9.69	9.60
RT	9.05	10.02	9.33	9.57	10.04	10.27	10.15
RF	8.81	8.90	9.12	9.11	**9.26**	9.81	9.70
SVR	8.73	8.84	9.16	9.34	9.51	9.70	9.62
ELM	**8.59**	8.66	9.48	9.36	9.45	9.66	9.65
FCDNN	8.66	8.69	**8.86**	**9.07**	9.35	**9.48**	**9.46**
Average	8.76	8.98	9.18	9.29	9.52	9.77	9.70
Improvement	2.77%	2.39%	0.22%	2.31%	2.56%	2.59%	1.62%

Table 7
Results (MAE) for Madrid database applying the threshold-based segmentation approach.

	+1 day	+2 days	+3 days	+4 days	+5 days	+6 days	+7 days
LR	**8.38**	8.96	9.17	**9.43**	9.47	**9.50**	**9.68**
RT	9.20	9.73	9.78	9.94	9.84	10.06	10.36
RF	8.66	9.08	9.42	**9.43**	9.48	9.73	10.04
SVR	8.44	8.99	9.18	**9.43**	9.47	9.52	9.79
ELM	8.48	**8.88**	**9.14**	10.97	9.54	9.63	11.11
FCDNN	8.47	8.93	9.24	9.45	**9.43**	9.54	9.89
Average	8.61	9.10	9.32	9.78	9.54	9.66	10.15
Improvement	3.15%	1.62%	1.89%	0.00%	2.15%	1.63%	-0.79%

number of visitors, a scenario where the base models tend to falter. It is worth remembering that this is precisely the idea behind the implementation of the continuous learning methodology.

3.5. Data segmentation

Results for the two data segmentation approaches introduced in this paper are reported and analyzed in this section. In addition, these methodologies allow an explainability analysis that enables the significance study of each driver for the different models assessed in both problems.

3.5.1. Threshold-based segmentation

Tables 6 and 7 show the results obtained after applying the threshold-based segmentation methodology on Pamplona and Madrid databases, respectively. Here, it is possible to observe the goodness of this approach in terms of MAE results, compared to the previous cases. The improvement percentage is computed in comparison to the continuous training case, showing that in almost all cases the average performance of the models is further improved when applying the threshold-based segmentation. Also, it is important to remark that for the cases where the average improvement is negative, the best performing model does improve when using the data segmentation approach.

Nonetheless, the primary advantage of this approach extends beyond merely improving numerical outcomes: it also enhances the capacity to provide explainability to the problem, by allowing an analysis of which predictor variables hold greater significance for the specific problem. In this threshold-based segmentation approach, an ensemble of models is employed, in such a way that a single prediction is computed when considering each threshold for segmenting the data, and all of the individual predictions are merged in an average ensemble to calculate the final prediction. These results are indicated in Tables 6 and 7. Now, the individual predictions considering all the thresholds taken into account are represented in Figs. 13 and 14 for Pamplona and Madrid cases, respectively. These figures correspond to the best performing model, ELM for Pamplona database and LR for Madrid database, and for the time prediction horizon of +1 day. Here, each vertical line corresponds to a threshold, and its color indicates the performance of the threshold-based prediction when using that specific threshold.

In this context, a comprehensive examination of variables and their associated thresholds has been conducted. For both cases, it can be observed that the variable that yields the better results is the day of the week in which the prediction is carried out. Specifically, the best results in terms of MAE are obtained when dividing the database in two distinct subsets, Mondays and Tuesdays in one, and Tuesday to Sunday in the other. Additionally, in the case of Pamplona database, meteorological variables appear to be also relevant. Notably, specific thresholds related to maximum and minimum temperatures yield optimal results. In particular, threshold at 273 K for minimum temperature and 293 K for maximum temperature obtain good predictive performance in the data segmentation approach, meaning that dividing the dataset for considering very cold days or extremely hot days separately helps the model improve the forecast of ED visits. Also, wind speed, msl and total precipitation present some relevant thresholds. Finally, when looking at the variables concerning the visits of previous days, the best threshold is found at the number of ED visits in the sixth previous day, when splitting the dataset using the number of 40 visits as threshold.

3.5.2. Cluster-based segmentation

Finally, this section reports the results obtained using the final proposed methodology, the cluster-based segmentation. In this study, a number of clusters between 2 and 10 has been considered, and for each case, predictions were generated. Tables 8 and 9 present the outcomes, which are derived by averaging the predictions across all these cases. These tables represent the MAE errors for Pamplona and Madrid databases, respectively.

In these tables, it becomes apparent that the utilization of this methodology results in a noteworthy improvement in the predictive capabilities of certain models, when compared to continuous training. Particularly, CART models: RT, and RF, exhibit a substantial increase in the quality of their predictions. However, in the case of neural networks models (ELM and FCDNN), this methodology does not lead to a significant improvement in prediction quality. Instead, the obtained results are a bit worse than those previously obtained. This can be explained by the fact that these models require a larger dataset to be effectively trained. When the database is divided into a larger number of clusters, the performance of these models tends to deteriorate. Conversely, in the prior scenario involving the threshold-based segmentation methodology, these models performed optimally due to the division of the database was performed into just two subsets, thereby maintaining a high volume of training data.

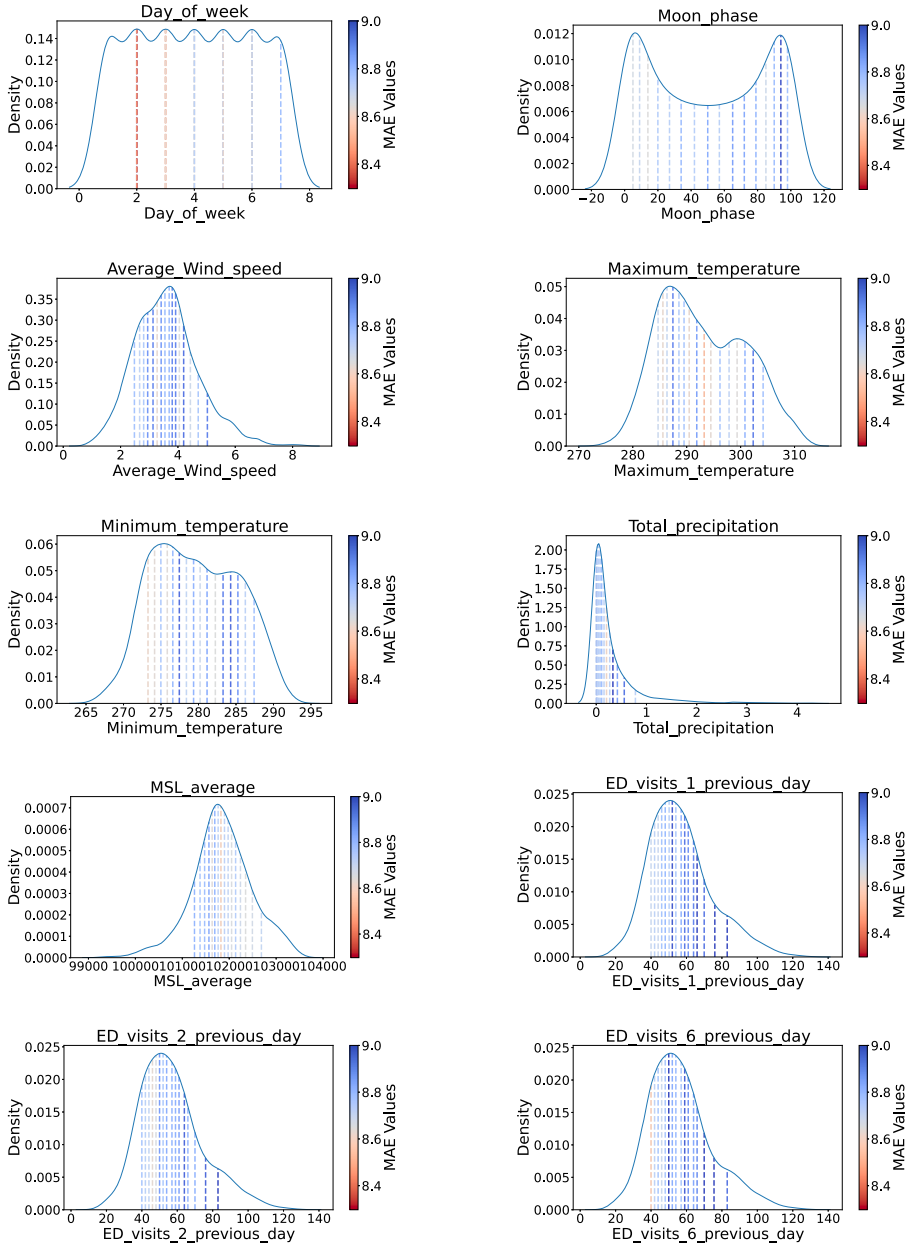

Fig. 13. Explainability analysis of the threshold-based segmentation approach. Vertical lines represent the different thresholds tested, and colors represent the prediction performance of each threshold, red colors indicate better performance in terms of MAE. The figure is extracted for the +1 day case of Pamplona database, and for the best performing model, ELM (For clarity, only the three most representative variables associated with previous ED visits have been included in the figure).

Table 8
Results (MAE) for Pamplona database applying the cluster-based segmentation approach.

	+1 day	+2 days	+3 days	+4 days	+5 days	+6 days	+7 days
LR	8.98	**8.83**	9.23	9.46	9.62	9.98	**9.72**
RT	9.17	9.66	9.57	9.73	9.76	10.03	10.27
RF	**8.85**	8.95	**9.13**	**9.21**	**9.27**	**9.65**	9.74
SVR	9.00	8.97	9.31	9.55	9.70	9.97	9.76
ELM	8.99	8.85	9.74	10.02	9.95	10.29	9.94
FCDNN	9.04	8.90	9.24	9.59	9.62	9.76	9.83
Average	9.01	9.03	9.37	9.59	9.65	9.95	9.88
Improvement	0.00%	1.85%	0.00%	-0.84%	1.23%	0.80%	-0.20%

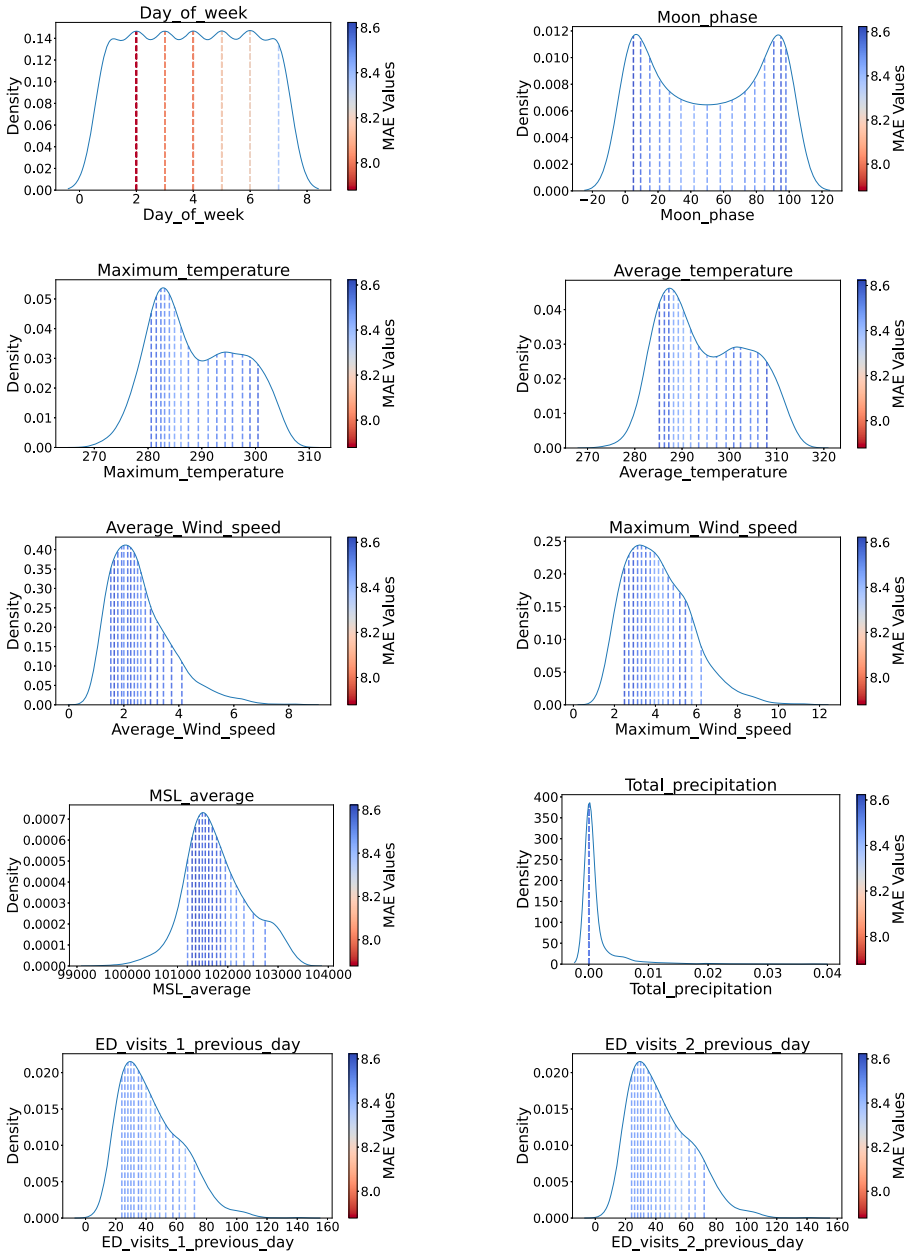

Fig. 14. Explainability analysis of the threshold-based segmentation approach. Vertical lines represent the different thresholds tested, and colors represent the prediction performance of each threshold, red colors indicate better performance in terms of MAE. The figure is extracted for the +1 day case of Madrid database, and for the best performing model, LR (For clarity, only the two most representative variables associated with previous ED visits have been included in the figure).

Table 9
Results (MAE) for Madrid database applying the cluster-based segmentation approach.

	+1 day	+2 days	+3 days	+4 days	+5 days	+6 days	+7 days
LR	8.40	9.16	9.49	9.71	9.73	**9.70**	**9.87**
RT	9.12	9.65	9.66	9.80	10.71	10.16	10.66
RF	8.51	**8.92**	**9.27**	**9.37**	**9.53**	**9.70**	10.08
SVR	**8.36**	9.27	9.55	9.76	9.69	9.71	9.95
ELM	9.55	9.44	10.05	11.99	10.05	11.04	10.39
FCDNN	8.53	9.07	9.46	9.65	9.64	9.74	10.15
Average	8.75	9.25	9.58	10.05	9.89	10.01	10.18
Improvement	1.57%	0.00%	-0.84%	-2.76%	-1.43%	-1.93%	-1.09%

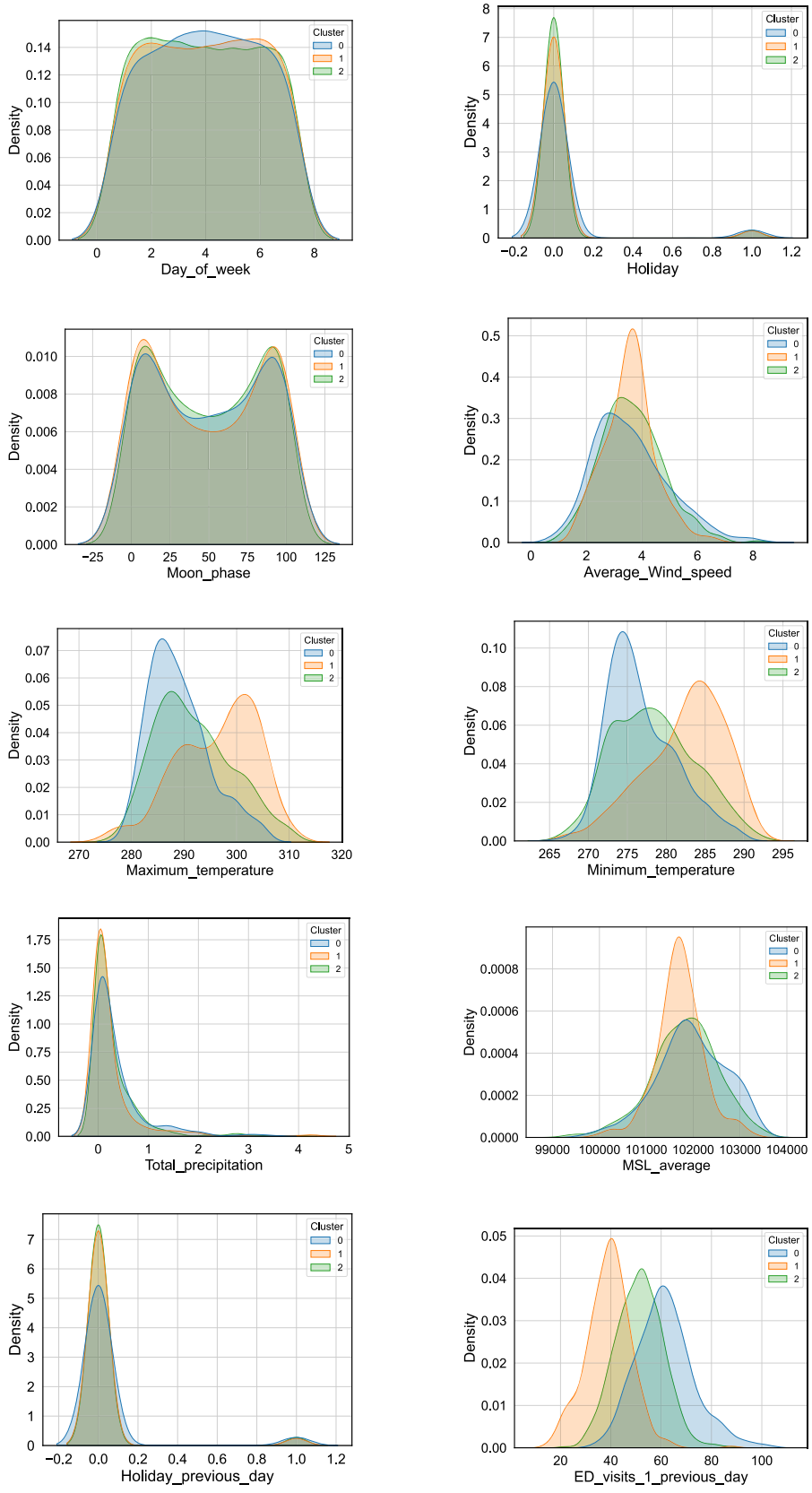

Fig. 15. Density distribution of each predictive variable across the different clusters, computed by setting the number of cluster to three, for Pamplona database.

Moreover, this methodology facilitates the examination of the acquired clusters, enabling a deeper understanding of which variables hold greater significance and exert dominance in the predictive problem. Figs. 15 and 16 aim to provide this analysis, in this case, the study was performed by configuring the number of clusters to be three, so that the interpretation of the figures is more accessible and intuitive.

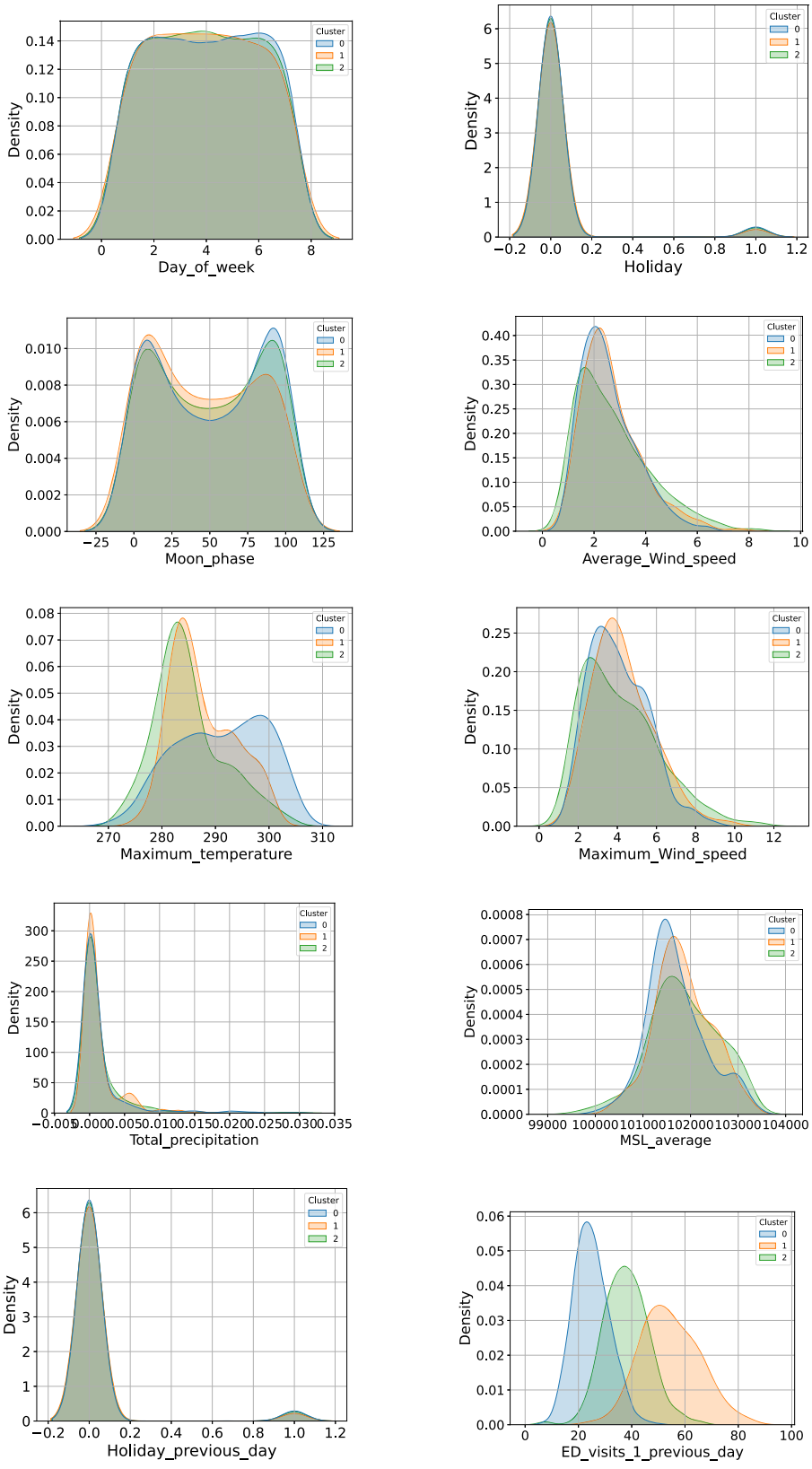

Fig. 16. Density distribution of each predictive variable across the different clusters, computed by setting the number of cluster to three, for Madrid database.

Each predictor variable is depicted in a subfigure, with different colors denoting the three calculated clusters. The X-axis reflects the range of the variable under investigation, while the figure illustrates the variable's distribution within each cluster. This distribution is represented by kernel density estimation, with the Y-axis indicating the kernel density values.

In this context, it is possible to discern that certain variables exhibit closely aligned distributions across the three clusters, whereas others

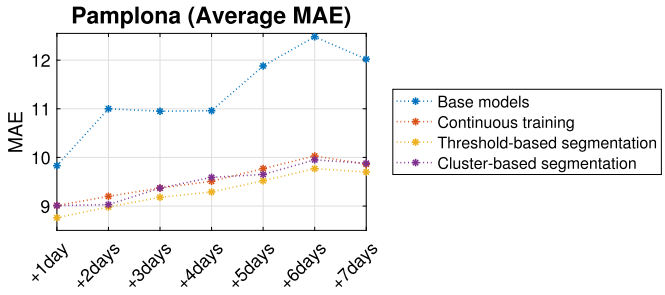

Fig. 17. Averaged performance evaluation of the six ML models, assessed in terms of MAE, across the four methodologies introduced in this paper. The analysis is conducted on Pamplona database and spans the seven distinct prediction time horizons.

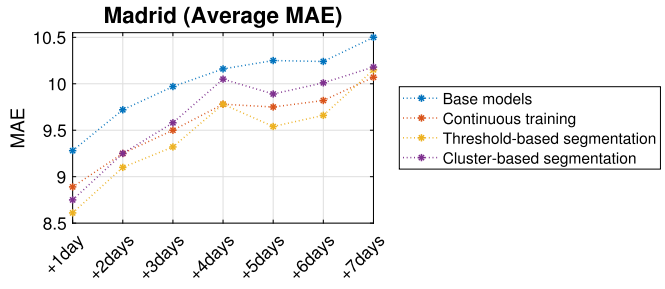

Fig. 18. Averaged performance evaluation of the six ML models, assessed in terms of MAE, across the four methodologies introduced in this paper. The analysis is conducted on Madrid database and spans the seven distinct prediction time horizons.

display markedly distinct distributions. This observation suggests that the first group of variables holds limited significance in the database clustering process, as their values do not notably influence the partitioning of the dataset. Conversely, the variables with divergent distributions play a pivotal role in governing the clustering procedure, emphasizing their greater weight and influence in the database.

Consequently, it can be deduced that the most distinctive variables in both databases are those related to temperature (specifically, maximum and minimum temperatures for Pamplona hospital, and maximum temperature for Madrid hospital) as well as the number of visitors on preceding days (only the variable referring to ED visits in the previous day has been represented for the shake of clarity, as the remaining ED visits variable exhibits very similar distributions). Therefore, this methodology suggests that special consideration should be directed towards these variables when attempting to forecast ED visits.

3.6. Comparison of results

In this section, a comprehensive examination and comparison of the results attained through the diverse methodologies employed in this paper, are presented.

First, a comparison among the performance of the different methodologies across the several prediction time-horizons considered is depicted in Figs. 17 and 18, considering the average of the six ML models used throughout the paper for Pamplona and Madrid hospitals, respectively. In this context, a notable improvement can be observed in the three methodologies that incorporate continuous training when compared to the base models, specially in Pamplona database when improvement rates rises from 8.34% to a 19.63%. Subsequently, albeit with a comparatively lower improvement ratio, it becomes apparent that the threshold-based segmentation methodology consistently outperforms continuous training for both databases in all prediction time-horizons except for the +7 days case of Madrid hospital. Regarding the outcomes achieved using the cluster-based segmentation methodology, it is observed that it yields a very similar performance to continuous

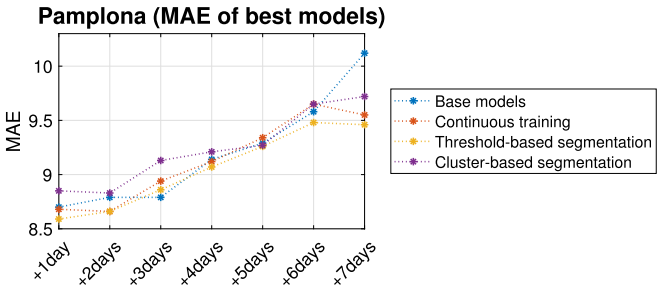

Fig. 19. Comparison of MAE results for the top-performing models acquired through each methodology for each time horizon on Pamplona database.

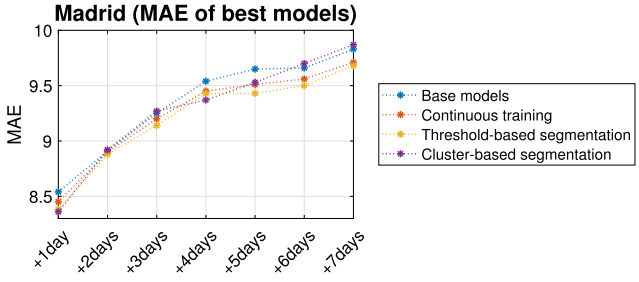

Fig. 20. Comparison of MAE results for the top-performing models acquired through each methodology for each time horizon on Madrid database.

training. In certain scenarios, it exhibits enhancements, while in others, there is a slight degradation. It is important to underscore that, with equivalent performance, this approach holds the added benefit of offering explanatory insights into the prediction problem.

Next, rather than depicting the average results obtained by the ML models, Figs. 19 and 20 illustrate the prediction errors of the top-performing models acquired through each methodology for each prediction time-horizon. These figures allow to analyze how, in the majority of cases (13 out of 14), the most favorable results are achieved through the explainable methodologies based on data segmentation approaches that have been proposed in this paper as the main contributions. When comparing the two databases, it can be noted that, for Pamplona dataset, the threshold-based segmentation strategy demonstrates significantly superior performance, dominating the best solutions in 6 out of 7 time horizons. On the other hand, in the case of the Madrid hospital, the cluster-based and threshold-based strategies exhibit much closer performance, with the cluster-based method achieving the best solutions in 3 out of 7 horizons and the threshold-based method leading in 4 out of 7.

Finally, this section concludes by presenting the temporal predictions made by the top performing models for one of the prediction time-horizons (+1 day) in both databases. Figs. 21 and 22 depict these graphs for Pamplona and Madrid hospitals, respectively, showing the high quality of the predictions provided by these models.

3.7. Implications and limitations

Obtaining a robust, high-performance, prediction system for hospital ED visits has a deep impact in the operation department of hospitals. Even accurate predictions within 1 day in advance prediction time-horizon would be enough to improve hospital operations, in terms of human resources and facilities such as operating rooms availability, etc. Longer prediction time-horizons would be even better in terms of hospital operations management. We have shown that increasing the prediction time-horizon is possible, and the worsen of the prediction results is controlled with the proposed approaches (continuous training, threshold-based and cluster-based segmentation), or at least better than applying the ML-based models on their own. Up to +7 days prediction

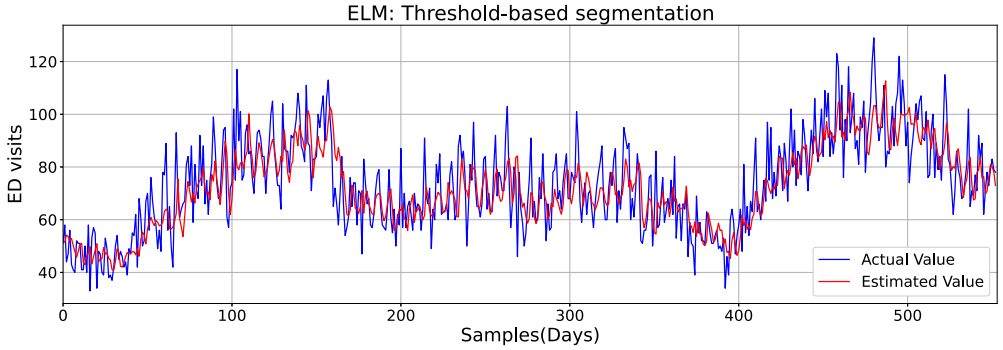

Fig. 21. Actual vs predicted time series values for the +1 day horizon. The prediction is computed using the best performing model for Pamplona database: ELM using the threshold-based segmentation.

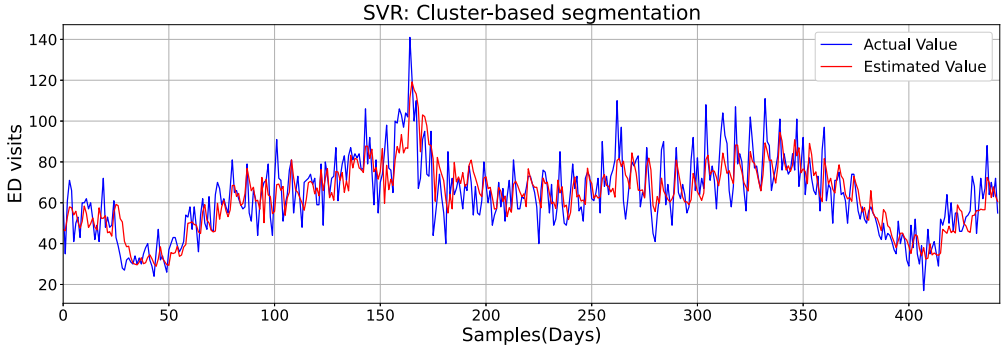

Fig. 22. Actual vs predicted time series values for the +1 day horizon. The prediction is computed using the best performing model for Madrid database: ELM using the threshold-based segmentation.

time-horizon has been analyzed, with results slightly worse than that for daily basis prediction. The threshold-based segmentation seems to be the most robust approach to implement the ML prediction in this particular problem, in both hospitals at Madrid and Pamplona. A detailed analysis of the input variables, how significant are they and what is their exact contribution to the ML prediction has been carried out. This analysis is important to improve the explainability of the results obtained, linked to the available input variables.

Another important implication of the current study is that it opens the possibility of spreading the proposed prediction system to forecast extreme events in hospital ED (large incoming of ED visits within a given period of time). Obtaining reasonable prediction of extreme events in the context of hospital ED would definitively improve the hospital operations, and would allow preparation against collapse of the department due to huge peaks of patients. In this case, the possible application of the proposed prediction system to extreme events depends on obtaining a significant amount of data to train the ML algorithms, which can be an issue in some cases, since in general there are not many extreme events cases in a normal operation of hospitals. In terms of ML algorithmic, extreme events prediction can be tackled as classification problems (see [74] in other application contexts), usually unbalanced in terms of the majority class (no extreme event) versus the minority class (existence of extreme event). Specific AI-based approaches focused on extreme events prediction could be applied to improve the prediction system in this particular (but very important) case.

Regarding limitations of the prediction system proposed, note that the problem of hospital ED visits forecast belongs to those prediction tasks related to with human-activity patterns [61] (other examples are human mobility and traffic patterns prediction problems, energy and power consumption or attendance to events, among others). Prediction problems associated with human activities are especially difficult to solve, because of their specific characteristics, and the scarce number of predictive (input) variables available to tackle these problems [61]. In general, this kind of problems the time tag (hour of the day, day of the week, etc.) in which the prediction is carried out is extremely important, together with meteorology predictive variables, which are usually associated with these problems. This is exactly the case of the hospital ED visits forecasting, where the input variables are also scarce and sometimes not very informative. Specific weather-related variables can partially improve the prediction, as we have discussed in the experiments carried out. In general, obtaining a good set of input variables to improve the performance of ML algorithms in this prediction task is a problem itself, which limits the performance of the final prediction system proposed.

4. Discussion and conclusions

Providing accurate and enough in-advance prediction of daily patient arrivals at ED, has emerged as a critical requirement in modern hospitals [75,76]. This is essential for ensuring the appropriate allocation of resources, encompassing staffing, equipment, and treatment facilities. A problem of forecasting ED visits using real data from two of important hospitals in Spain has been tackled in this paper, using seven different prediction time-horizons covering from 1 to 7 days. This forecasting task has been carried out using ML-based algorithms, and with a strong emphasis on providing explainability to the problem, so that it is possible to understand which variables govern the problem and are pivotal for obtaining precise predictions. Although ML has been applied extensively to solve ED related forecast problems [35,37–39,42], there are not many studies which have focused on the transparency and explainability of the problem.

The main novelties and contributions presented in this work consist of the introduction of two explainable prediction methodologies, based on data segmentation and continuous training. These methodologies facilitate a significant enhancement in predictive performance, while concurrently offering valuable insights into the prediction pro-

cess. They enable the extraction of transparent conclusions regarding the relative significance of predictive variables in the forecast process. The proposed approaches use two different criteria for segmenting the database: First, a threshold-based strategy is employed, which involves segmenting data based on specific predictor variable values, and training separate ML with each subset. Second, a cluster-based ML ensemble method is proposed. In this scenario, a clustering algorithm is applied to the training dataset, followed by the training of ML models for each cluster. When predicting a new sample, we identify the nearest cluster for that sample and utilize the corresponding ML model to make predictions. The outcomes derived from these strategies have demonstrated a notable improvement over the initial results obtained with ML base models. The proposed approaches emerge as top-performing models, in terms of prediction metrics, across the seven different time-horizons assessed in Pamplona database and in six out of seven time-horizons in Madrid database.

Regarding the performance of the specific six ML models assessed in the paper, a consistent and balanced performance is observed for all of them in Madrid database, as evidenced by the results across the seven tested prediction time-horizons. Remarkably, five different models (2 LR, 1 SVR, 1 RF, 1 FCDNN and 2 ELM) yield the best results. On the other hand, the prediction in Pamplona hospital is dominated by neural networks models, in four of the seven cases the best model corresponds to the FCDNNN, one to the ELM, one to the RF and one is shared by ELM and FCDNNN.

An important aspect addressed in this research is related to the proposition of transparent solutions, that yield interpretable outcomes for the problem at hand. In this regard, the following conclusions can be drawn from the first assessed methodology, the threshold-based segmentation: (1) Dividing the database according to the day of the week allows obtaining the best results in both databases. More precisely, partitioning the data into two subsets, with Mondays and Tuesdays on one side, and the remaining weekdays on the other, results in the most accurate predictions; (2) Furthermore, in the case of Pamplona database, meteorological variables exhibit notable significance. Particularly, specific thresholds associated with maximum and minimum temperatures have been found to yield optimal results. This implies that dividing the dataset to differentiate between very cold and warm days substantially aids in improving the forecast of ED visits. Additionally, variables like wind speed, mean sea level pressure (msl), and total precipitation show noteworthy thresholds; (3) Finally, regarding variables related to visits on previous days, the most effective threshold is identified at 40 visits when splitting the dataset based on the count of ED visits from the sixth previous day. This information holds significant relevance in addressing the prediction problem, emphasizing the importance of not only relying on the ML model's output, but rather conducting further analysis to account for these considerations.

In addition, the analysis of the cluster-based segmentation methodology reveals the following insights: (1) Some predictive variables hold limited significance in the database clustering process, as their values do not notably influence the partitioning of the dataset. However, others play a pivotal role in governing the clustering procedure, emphasizing their greater weight and influence in the database; (2) Variables related to temperature hold significant relevance in both databases when performing the training data segmentation into various subsets; (3) Variables referring to the number of visitors on preceding days also play a pivotal role in the process.

When we combine the insights derived from both methodologies, we can presume that special consideration should be directed towards specific variables when aiming to predict ED visits. These crucial variables include temperature, prior days' visitor count, and day of the week. Additionally, other meteorological variables also exhibit relevance, albeit to a lesser extent. These findings are consistent with those observed in previous literature, which give high importance to weather-related variables as temperature and precipitation, as well as volume of patients in previous days [58,40]. In [46] SHAP analysis reveals that day of the week, mean number of visits on the four previous on duty days and daily maximum temperature are the most important predictive variables. Also, partitioning the data into two subsets according to the day of the week with Mondays and Tuesdays on one side and the remaining weekdays on the other aligns with previous literature, which find that Monday is the busiest day of the week [31,77,47]. Regarding the temperature, we have observed a threshold at 273 K (0 ºC) for minimum temperature and 293 K (20 ºC) for maximum obtained good predictive performance. This agrees with other works in the literature claiming that days with a maximum temperature above 297 K (24 ºC) have been found to have an increased probability for larger numbers of patient ED visits while [46,78,41].

Future lines of research will be focused on different aspects of the prediction model developed: First, the development of hybrid models that combine the strengths of the explainable data segmentation methodologies presented with the powerful capabilities of more complex deep learning regression models. Additionally, future research will delve into refining the selection of optimal thresholds and variables to improve forecasting precision, possibly using evolutionary computation instead of applying an average ensemble approach. Also, a more complex problem can be defined, involving the distinction between various patient categories and an extension of the prediction-time horizons of the forecasting problem. Finally, the adaptation of the prediction system to extreme events (huge visits number to the hospital's ED in a short period of time), will be very useful to improve the operations, management and logistics of the hospital.

CRediT authorship contribution statement

C. Peláez-Rodríguez: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft. **R. Torres-López:** Conceptualization, Data curation, Formal analysis, Writing – original draft. **J. Pérez-Aracil:** Data curation, Investigation, Methodology, Software. **N. López-Laguna:** Resources, Supervision, Visualization, Writing – review & editing. **S. Sánchez-Rodríguez:** Resources, Supervision, Validation, Writing – review & editing. **S. Salcedo-Sanz:** Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

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

This research has been partially supported by the project PID2020-115454GB-C21 of the Spanish Ministry of Science and Innovation (MICINN). This research has been partially funded by the Autonomous Community of Madrid through ELLIS consortium.

References

[1] S. Di Somma, L. Paladino, L. Vaughan, I. Lalle, L. Magrini, M. Magnanti, Overcrowding in emergency department: an international issue, Intern. Emerg. Med. 10 (2015) 171–175.
[2] L.I. Horwitz, J. Green, E.H. Bradley, Us emergency department performance on wait time and length of visit, Ann. Emerg. Med. 55 (2) (2010) 133–141.
[3] H. Lau, A. Dadich, D. Nakandala, H. Evans, L. Zhao, Development of a cost-optimization model to reduce bottlenecks: a health service case study, Expert Syst. 35 (6) (2018) e12294.
[4] H. Boerner, A 'durable opportunity': ED overcrowding in the ACA ERA, Phys. Leadersh. J. 3 (3) (2016) 32.
[5] R. Forero, S. McCarthy, K. Hillman, Access block and emergency department overcrowding, in: Annual Update in Intensive Care and Emergency Medicine, 2011, pp. 720–728.

[6] A.J. Forster, I. Stiell, G. Wells, A.J. Lee, C. Van Walraven, The effect of hospital occupancy on emergency department length of stay and patient disposition, Acad. Emerg. Med. 10 (2) (2003) 127–133.
[7] P. Cremonesi, E. di Bella, M. Montefiori, L. Persico, The robustness and effectiveness of the triage system at times of overcrowding and the extra costs due to inappropriate use of emergency departments, Appl. Health Econ. Health Policy 13 (2015) 507–514.
[8] J.M. Pines, J.A. Hilton, E.J. Weber, A.J. Alkemade, H. Al Shabanah, P.D. Anderson, M. Bernhard, A. Bertini, A. Gries, S. Ferrandiz, et al., International perspectives on emergency department crowding, Acad. Emerg. Med. 18 (12) (2011) 1358–1370.
[9] R.E. Kheirbek, S. Beygi, M. Zargoush, F. Alemi, A.W. Smith, R.D. Fletcher, P.N. Seton, B.A. Hawkins, Causal analysis of emergency department delays, Qual. Manag. Healthc. 24 (3) (2015) 162–166.
[10] G. Bouzillé, C. Poirier, B. Campillo-Gimenez, M.-L. Aubert, M. Chabot, E. Chazard, A. Lavenu, M. Cuggia, Leveraging hospital big data to monitor flu epidemics, Comput. Methods Programs Biomed. 154 (2018) 153–160.
[11] X. Zhao, J.W. Lai, A.F.W. Ho, N. Liu, M.E.H. Ong, K.H. Cheong, Predicting hospital emergency department visits with deep learning approaches, Biocybern. Biomed. Eng. 42 (3) (2022) 1051–1065.
[12] K.L. Khatri, L.S. Tamil, Early detection of peak demand days of chronic respiratory diseases emergency department visits using artificial neural networks, IEEE J. Biomed. Health Inform. 22 (1) (2017) 285–290.
[13] A. Abidova, P.A.d. Silva, S. Moreira, Predictors of patient satisfaction and the perceived quality of healthcare in an emergency department in Portugal, West. J. Emerg. Med. (2020) 1–12.
[14] R.W. Derlet, J.R. Richards, Overcrowding in the nation's emergency departments: complex causes and disturbing effects, Ann. Emerg. Med. 35 (1) (2000) 63–68.
[15] M. Wargon, B. Guidet, T. Hoang, G. Hejblum, A systematic review of models for forecasting the number of emergency department visits, J. Emerg. Med. 26 (6) (2009) 395–399.
[16] Y.-H. Hu, C.-T. Tai, S.C.-C. Chen, H.-W. Lee, S.-F. Sung, Predicting return visits to the emergency department for pediatric patients: applying supervised learning techniques to the Taiwan national health insurance research database, Comput. Methods Programs Biomed. 144 (2017) 105–112.
[17] M. Gul, E. Celik, An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments, Health Syst. 9 (4) (2020) 263–284.
[18] I. Marcilio, S. Hajat, N. Gouveia, Forecasting daily emergency department visits using calendar variables and ambient temperature readings, Acad. Emerg. Med. 20 (8) (2013) 769–777.
[19] A.F.W. Ho, B.Z.Y.S. To, J.M. Koh, K.H. Cheong, Forecasting hospital emergency department patient volume using Internet search data, IEEE Access 7 (2019) 93387–93395.
[20] Y. Sun, B.H. Heng, S.Y. Tay, E. Seow, Predicting hospital admissions at emergency department triage using routine administrative data, Acad. Emerg. Med. 18 (8) (2011) 844–850.
[21] C. Chatfield, H. Xing, The Analysis of Time Series: an Introduction with R, CRC Press, 2019.
[22] W.-C. Juang, S.-J. Huang, F.-D. Huang, P.-W. Cheng, S.-R. Wann, Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in southern Taiwan, BMJ Open 7 (11) (2017) e018628.
[23] P. Milner, Ten-year follow-up of arima forecasts of attendances at accident and emergency departments in the Trent region, Stat. Med. 16 (18) (1997) 2117–2125.
[24] L.M. Schweigler, J.S. Desmond, M.L. McCarthy, K.J. Bukowski, E.L. Ionides, J.G. Younger, Forecasting models of emergency department crowding, Acad. Emerg. Med. 16 (4) (2009) 301–308.
[25] Y. Sun, B.H. Heng, Y.T. Seow, E. Seow, Forecasting daily attendances at an emergency department to aid resource planning, BMC Emerg. Med. 9 (1) (2009) 1–9.
[26] S.S. Jones, A. Thomas, R.S. Evans, S.J. Welch, P.J. Haug, G.L. Snow, Forecasting daily patient volumes in the emergency department, Acad. Emerg. Med. 15 (2) (2008) 159–170.
[27] J. Bergs, P. Heerinckx, S. Verelst, Knowing what to expect, forecasting monthly emergency department visits: a time-series analysis, Int. Emerg. Nurs. 22 (2) (2014) 112–115.
[28] M. Butler, H. Gu, T. Kenney, S. Campbell, P017: does a busy day predict another busy day? A time-series analysis of multi-centre emergency department volumes, Can. J. Emerg. Med. 18 (S1) (2016) S83–S84.
[29] R. Calegari, F.S. Fogliatto, F.R. Lucini, J. Neyeloff, R.S. Kuchenbecker, B.D. Schaan, Forecasting daily volume and acuity of patients in the emergency department, Comput. Math. Methods Med. 2016 (2016).
[30] H.J. Kam, J.O. Sung, R.W. Park, Prediction of daily patient numbers for a regional emergency medical center using time series analysis, Healthc. Inform. Res. 16 (3) (2010) 158–165.
[31] Q. Cheng, N.T. Argon, C.S. Evans, Y. Liu, T.F. Platts-Mills, S. Ziya, Forecasting emergency department hourly occupancy using time series analysis, Am. J. Emerg. Med. 48 (2021) 177–182.
[32] S.S. Jones, R.S. Evans, T.L. Allen, A. Thomas, P.J. Haug, S.J. Welch, G.L. Snow, A multivariate time series approach to modeling and forecasting demand in the emergency department, J. Biomed. Inform. 42 (1) (2009) 123–139.
[33] N.R. Hoot, L.J. LeBlanc, I. Jones, S.R. Levin, C. Zhou, C.S. Gadd, D. Aronsky, Forecasting emergency department crowding: a discrete event simulation, Ann. Emerg. Med. 52 (2) (2008) 116–125.
[34] N.R. Hoot, L.J. LeBlanc, I. Jones, S.R. Levin, C. Zhou, C.S. Gadd, D. Aronsky, Forecasting emergency department crowding: a prospective, real-time evaluation, J. Am. Med. Inform. Assoc. 16 (3) (2009) 338–345.
[35] O.H. Salman, Z. Taha, M.Q. Alsabah, Y.S. Hussein, A.S. Mohammed, M. Aal-Nouman, A review on utilizing machine learning technology in the fields of electronic emergency triage and patient priority systems in telemedicine: coherent taxonomy, motivations, open research challenges and recommendations for intelligent future work, Comput. Methods Programs Biomed. 209 (2021) 106357.
[36] Y.-H. Kuo, N.B. Chan, J.M. Leung, H. Meng, A.M.-C. So, K.K. Tsoi, C.A. Graham, An integrated approach of machine learning and systems thinking for waiting time prediction in an emergency department, Int. J. Med. Inform. 139 (2020) 104143.
[37] M. Xu, T.-C. Wong, K.-S. Chin, Modeling daily patient arrivals at emergency department and quantifying the relative importance of contributing variables using artificial neural network, Decis. Support Syst. 54 (3) (2013) 1488–1498.
[38] N.B. Menke, N. Caputo, R. Fraser, J. Haber, C. Shields, M.N. Menke, A retrospective analysis of the utility of an artificial neural network to predict ed volume, Am. J. Emerg. Med. 32 (6) (2014) 614–617.
[39] W. Whitt, X. Zhang, Forecasting arrivals and occupancy levels in an emergency department, Oper. Res. Heal. Care 21 (2019) 1–18.
[40] P. Aboagye-Sarfo, Q. Mai, F.M. Sanfilippo, D.B. Preen, L.M. Stewart, D.M. Fatovich, A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in western Australia, J. Biomed. Inform. 57 (2015) 62–73.
[41] V.K. Sudarshan, M. Brabrand, T.M. Range, U.K. Wiil, Performance evaluation of emergency department patient arrivals forecasting models by including meteorological and calendar information: a comparative study, Comput. Biol. Med. 135 (2021) 104541.
[42] G. Gafni-Pappas, M. Khan, Predicting daily emergency department visits using machine learning could increase accuracy, Am. J. Emerg. Med. 65 (2023) 5–11.
[43] M.A. Vollmer, B. Glampson, T. Mellan, S. Mishra, L. Mercuri, C. Costello, R. Klaber, G. Cooke, S. Flaxman, S. Bhatt, A unified machine learning approach to time series forecasting applied to demand at emergency departments, BMC Emerg. Med. 21 (1) (2021) 1–14.
[44] A. Zlotnik, A. Gallardo-Antolin, M.C. Alfaro, M.C.P. Pérez, J.M.M. Martínez, et al., Emergency department visit forecasting and dynamic nursing staff allocation using machine learning techniques with readily available open-source software, Comput. Inf. Nurs. 33 (8) (2015) 368–377.
[45] Y. Zhang, L. Luo, J. Yang, D. Liu, R. Kong, Y. Feng, A hybrid arima-svr approach for forecasting emergency patient flow, J. Ambient Intell. Humaniz. Comput. 10 (2019) 3315–3323.
[46] S. Petsis, A. Karamanou, E. Kalampokis, K. Tarabanis, Forecasting and explaining emergency department visits in a public hospital, J. Intell. Inf. Syst. 59 (2) (2022) 479–500.
[47] C.N. Rocha, F. Rodrigues, Forecasting emergency department admissions, Intell. Data Anal. 25 (6) (2021) 1579–1601.
[48] F. Xie, J. Zhou, J.W. Lee, M. Tan, S. Li, L.S. Rajnthern, M.L. Chee, B. Chakraborty, A.-K.I. Wong, A. Dagan, et al., Benchmarking emergency department prediction models with machine learning and public electronic health records, Sci. Data 9 (1) (2022) 658.
[49] F. Harrou, A. Dairi, F. Kadri, Y. Sun, Forecasting emergency department overcrowding: a deep learning framework, Chaos Solitons Fractals 139 (2020) 110247.
[50] F. Kadri, K. Abdennbi, Rnn-based deep-learning approach to forecasting hospital system demands: application to an emergency department, Int. J. Data Sci. 5 (1) (2020) 1–25.
[51] F. Kadri, M. Baraoui, I. Nouaouri, An lstm-based deep learning approach with application to predicting hospital emergency department admissions, in: 2019 International Conference on Industrial Engineering and Systems Management (IESM), IEEE, 2019, pp. 1–6.
[52] T. Chen, All versus one: an empirical comparison on retrained and incremental machine learning for modeling performance of adaptable software, in: 2019 IEEE/ACM 14th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS), IEEE, 2019, pp. 157–168.
[53] I. Prapas, B. Derakhshan, A.R. Mahdiraji, V. Markl, Continuous training and deployment of deep learning models, Datenbank Spektrum 21 (3) (2021) 203–212.
[54] M. Barque, S. Martin, J.E.N. Vianin, D. Genoud, D. Wannier, Improving wind power prediction with retraining machine learning algorithms, in: 2018 International Workshop on Big Data and Information Security (IWBIS), IEEE, 2018, pp. 43–48.
[55] J. Treboux, R. Ingold, D. Genoud, Towards retraining of machine learning algorithms: an efficiency analysis applied to smart agriculture, in: 2020 Global Internet of Things Summit (GIoTS), IEEE, 2020, pp. 1–6.
[56] D. Nallaperuma, R. Nawaratne, T. Bandaragoda, A. Adikari, S. Nguyen, T. Kempitiya, D. De Silva, D. Alahakoon, D. Pothuhera, Online incremental machine learning platform for big data-driven smart traffic management, IEEE Trans. Intell. Transp. Syst. 20 (12) (2019) 4679–4690.
[57] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai, Inf. Fusion 58 (2020) 82–115.
[58] A.K. Diehl, M.D. Morris, S.A. Mannis, Use of calendar and weather data to predict walk-in attendance, South. Med. J. 74 (6) (1981) 709–712.

[59] T. Susnjak, P. Maddigan, Forecasting patient flows with pandemic induced concept drift using explainable machine learning, EPJ Data Sci. 12 (1) (2023) 11.
[60] C. Peláez-Rodríguez, J. Pérez-Aracil, D. Fister, R. Torres-López, S. Salcedo-Sanz, Bike sharing and cable car demand forecasting using machine learning and deep learning multivariate time series approaches, Expert Syst. Appl. (2023) 122264, https://doi.org/10.1016/j.eswa.2023.122264, https://www.sciencedirect.com/science/article/pii/S0957417423027665.
[61] R. Torres-López, D. Casillas-Pérez, J. Pérez-Aracil, L. Cornejo-Bueno, E. Alexandre, S. Salcedo-Sanz, Analysis of machine learning approaches' performance in prediction problems with human activity patterns, Mathematics 10 (13) (2022) 2187.
[62] Y. Ren, L. Zhang, P.N. Suganthan, Ensemble classification and regression-recent developments, applications and future directions, IEEE Comput. Intell. Mag. 11 (1) (2016) 41–53.
[63] H. Chen, Cluster-based ensemble learning for wind power modeling from meteorological wind data, Renew. Sustain. Energy Rev. 167 (2022) 112652.
[64] S. Tasnim, A. Rahman, A.M.T. Oo, M.E. Haque, Wind power prediction using cluster based ensemble regression, Int. J. Comput. Intell. Appl. 16 (04) (2017) 1750026.
[65] T. Jilani, G. Housley, G. Figueredo, P.-S. Tang, J. Hatton, D. Shaw, Short and long term predictions of hospital emergency department attendances, Int. J. Med. Inform. 129 (2019) 167–174.
[66] A. Rivera, J.C. Muñoz, M. Pérez-Goody, B.S. de San Pedro, F. Charte, D. Elizondo, C. Rodríguez, M. Abolafia, A. Perea, M. Del Jesus, Xaire: an ensemble-based methodology for determining the relative importance of variables in regression tasks. Application to a hospital emergency department, Artif. Intell. Med. 137 (2023) 102494.
[67] N.R. Draper, H. Smith, Applied Regression Analysis, vol. 326, John Wiley & Sons, 1998.
[68] W.-Y. Loh, Classification and regression trees, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1 (1) (2011) 14–23.
[69] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[70] W.S. Noble, What is a support vector machine?, Nat. Biotechnol. 24 (12) (2006) 1565–1567.
[71] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (2006) 489–501, https://doi.org/10.1016/j.neucom.2005.12.126.
[72] M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, M. Hasan, B.C. Van Essen, A.A. Awwal, V.K. Asari, A state-of-the-art survey on deep learning theory and architectures, Electronics 8 (3) (2019) 292.
[73] D. Steinley, K-means clustering: a half-century synthesis, Br. J. Math. Stat. Psychol. 59 (1) (2006) 1–34.
[74] C. Peláez-Rodríguez, J. Pérez-Aracil, D. Fister, L. Prieto-Godino, R. Deo, S. Salcedo-Sanz, A hierarchical classification/regression algorithm for improving extreme wind speed events prediction, Renew. Energy 201 (2022) 157–178.
[75] V.J. Chase, A.E. Cohn, T.A. Peterson, M.S. Lavieri, Predicting emergency department volume using forecasting methods to create a "surge response" for noncrisis events, Acad. Emerg. Med. 19 (5) (2012) 569–576.
[76] S.J. Littig, M.W. Isken, Short term hospital occupancy prediction, Health Care Manage. Sci. 10 (2007) 47–66.
[77] N.S. Erkamp, D.H. van Dalen, E. de Vries, Predicting emergency department visits in a large teaching hospital, Int. J. Emerg. Med. 14 (1) (2021) 1–11.
[78] K.I. Duwalage, E. Burkett, G. White, A. Wong, M.H. Thompson, Forecasting daily counts of patient presentations in Australian emergency departments using statistical models with time-varying predictors, Emerg. Med. Australasia 32 (4) (2020) 618–625.