Peter the Great St. Petersburg Polytechnic University
Institute of computer science and cybersecurity
Higher school of artificial intelligence

An explainable ML approach for hospital ED visits forecasting using continuous training and multi-model regression

Student: Salimli A. Gr. 5130201/20102
Supervisor: Ph.D in Technical Science, D.E Motorin

Autumn 2024

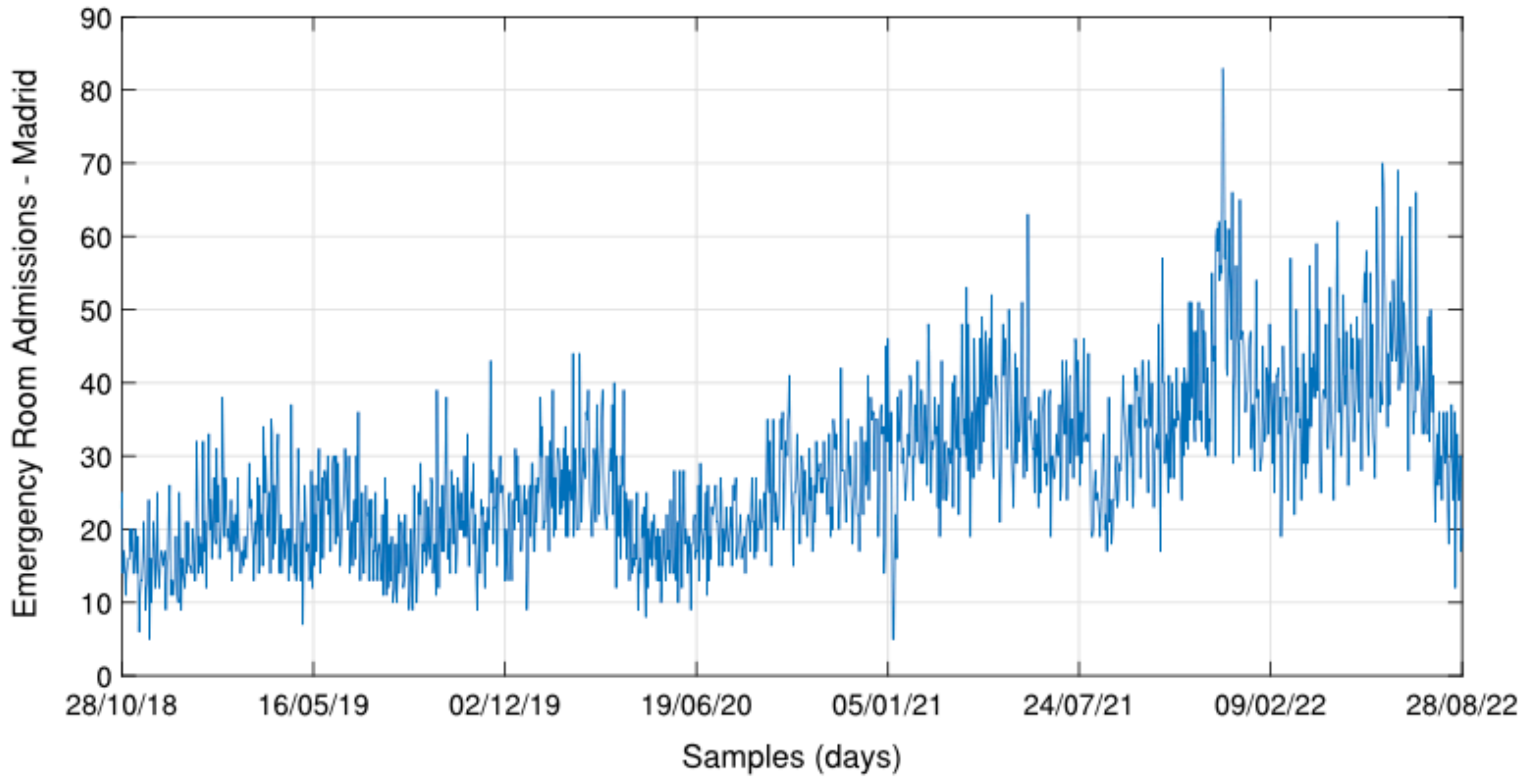
Introduction

<u>Significance</u>:

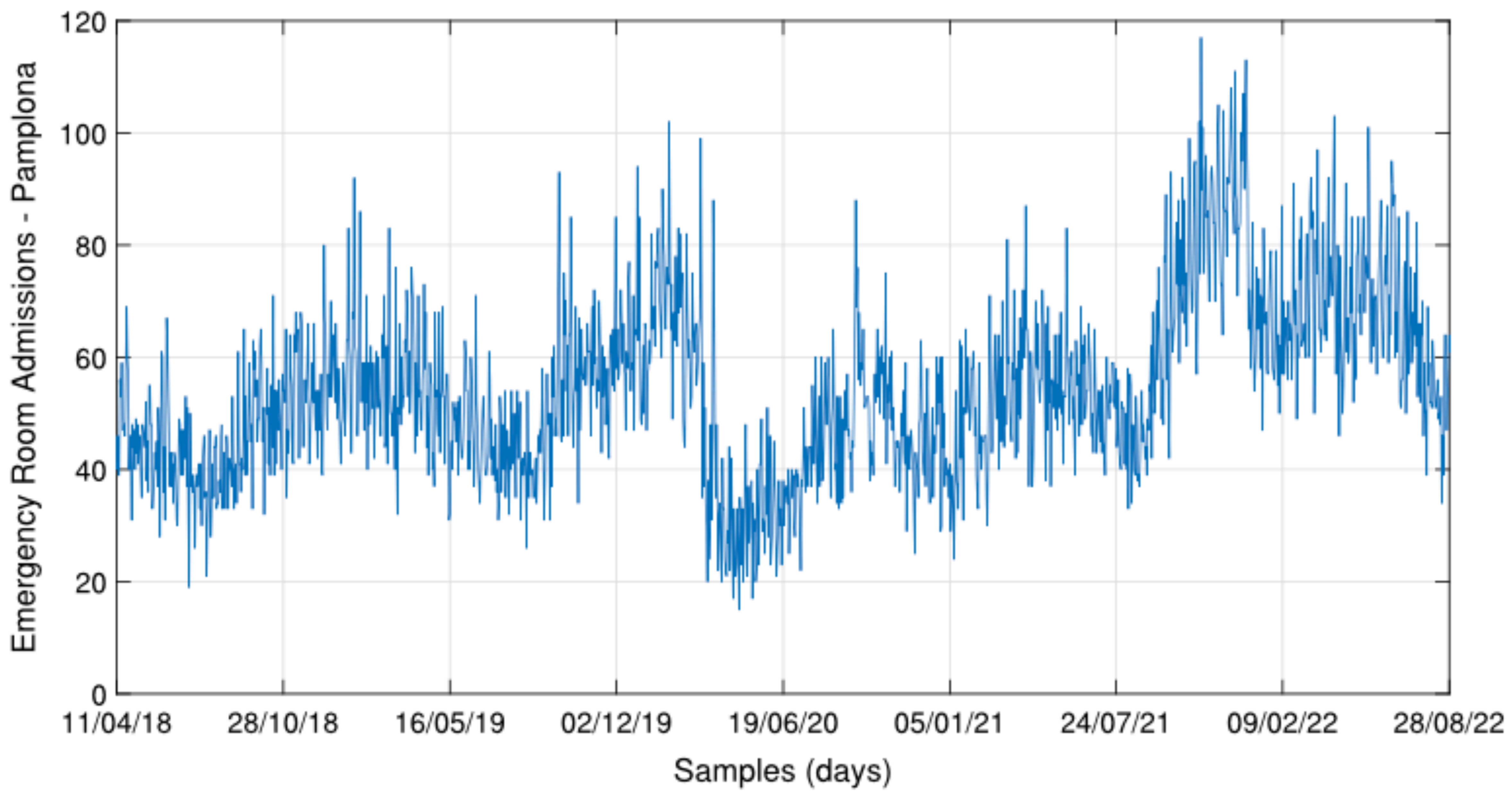
Explainability and Continuous Model Adaptation in Forecasting Emergency Department Visits

<u>Area of expertise</u>:

- Forecasting emergency department (ED) visits
- Machine learning methods
- Continuous training



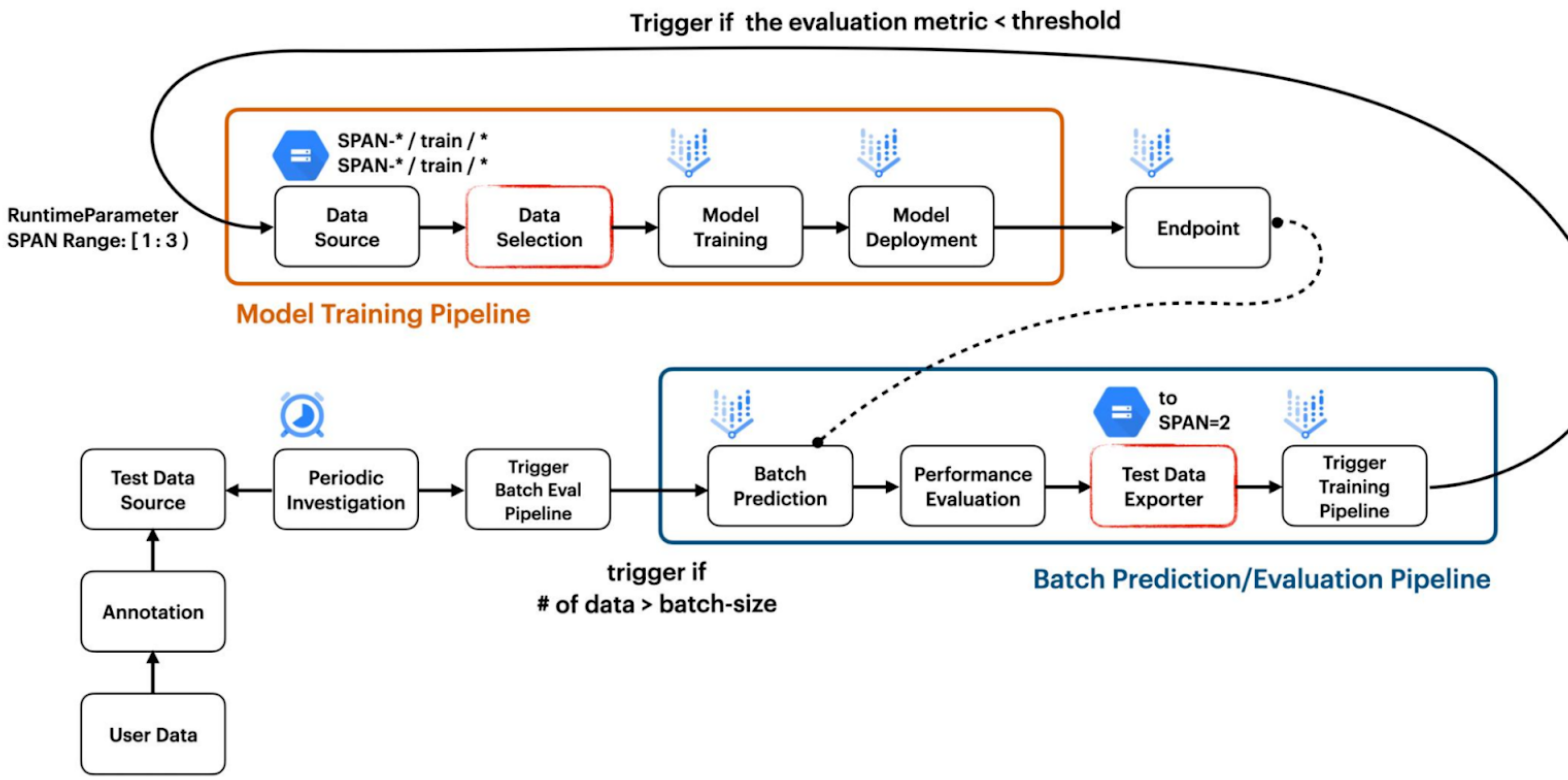
Graph of visits by years Madrid



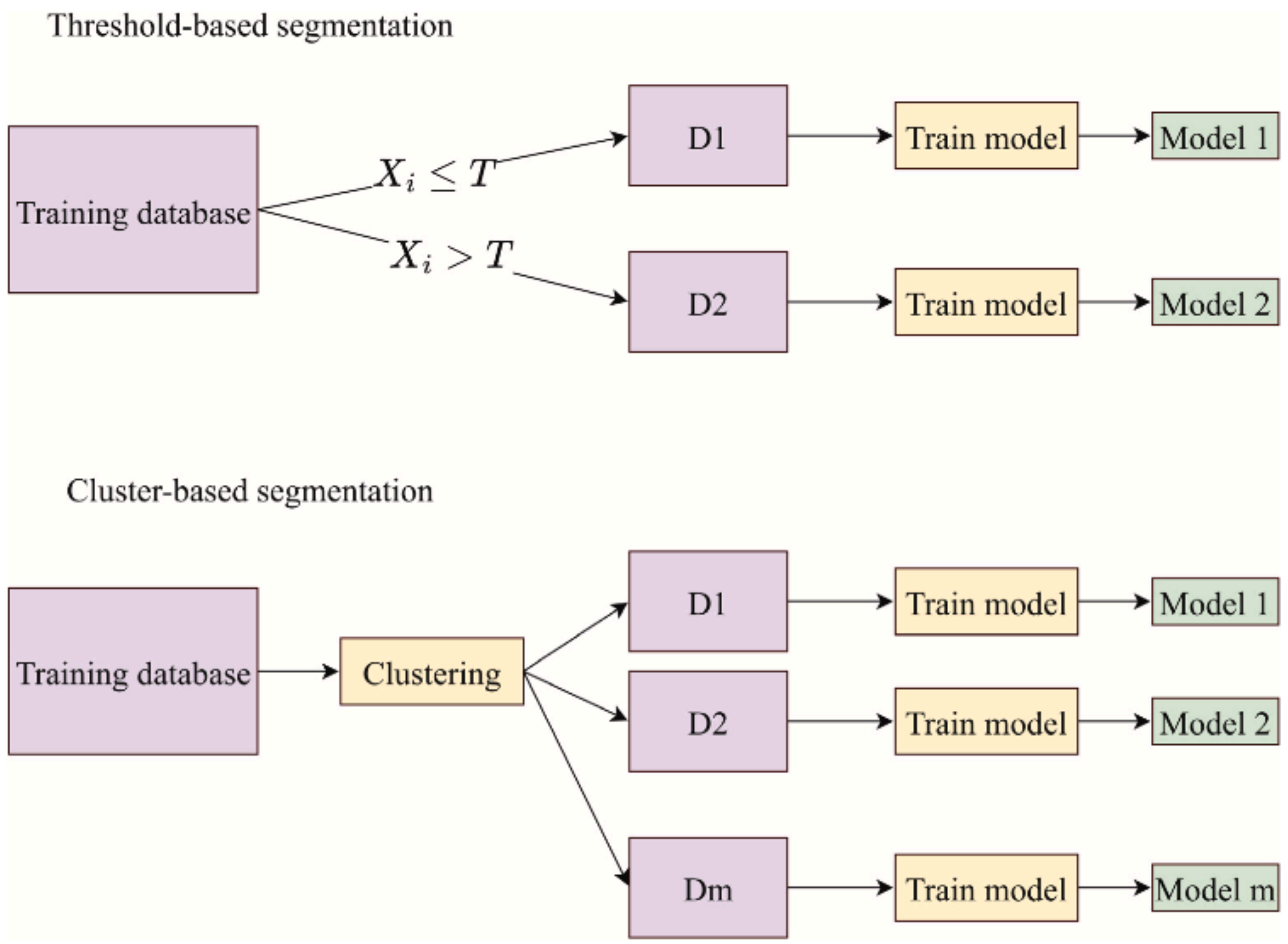
Graph of visits by years Pamplona

Problem statement

- Resource allocation
- Waiting time
- Old and classic ML methods
- Real-time data



Continuous Adaptation for Machine Learning System to Data Changes (TF)



Methods

Datasets

Real datas from two branches of hospital «Clínica Universidad de Navarra»:

- Pamplona (Navarre)
- Madrid (est. 2018)

Date	Day_of_week	Holiday	Moon_Phase	Average_Temp	Max_temp	Average_wind	Max_wind	Average_mslp	Total_precipitation	Holiday_prev	ED_1
08/09/2018	6	0	5	295.01	299.3	2.56	3.96	101703.64	0.00281	0.0	12.0
09/09/2018	7	0	1	295.68	302.16	1.26	3.03	101901.04	0.00012	0.0	12.0
10/09/2018	1	0	0	296.24	302.14	2.44	3.57	102190.41	0.00014	0.0	16.0
11/09/2018	2	0	1	296.4	303.19	2.36	3.57	102223.68	0.00139	0.0	21.0
12/09/2018	3	0	4	297.18	303.88	1.39	2.95	102153.5	1e-05	0.0	14.0
13/09/2018	4	0	10	299.03	306.49	1.14	1.8	101790.55	0.0	0.0	14.0
14/09/2018	5	0	17	298.79	305.69	1.67	2.79	101689.64	0.00013	0.0	17.0
15/09/2018	6	0	25	297.6	303.66	2.26	4.24	101878.94	0.00019	0.0	19.0
16/09/2018	7	0	35	297.1	304.5	0.99	1.6	101960.37	0.0	0.0	17.0
17/09/2018	1	0	46	296.91	304.45	2.54	4.4	101674.67	0.00428	0.0	19.0
18/09/2018	2	0	56	295.57	301.54	1.48	3.08	101663.15	0.00026	0.0	22.0
19/09/2018	3	0	66	297.15	304.5	1.21	2.13	101951.79	0.0	0.0	15.0
20/09/2018	4	0	76	297.6	304.83	1.54	2.41	101878.47	0.0	0.0	26.0
21/09/2018	5	0	84	298.13	305.4	1.42	2.62	101812.55	0.0	0.0	18.0
22/09/2018	6	0	91	298.76	306.65	1.93	2.71	102134.65	0.0	0.0	18.0
23/09/2018	7	0	96	299.17	307.51	1.93	3.18	102160.0	0.0	0.0	16.0
24/09/2018	1	0	99	299.19	306.35	3.03	4.54	102104.86	0.0	0.0	23.0

Part of dataset. Example (Madrid)

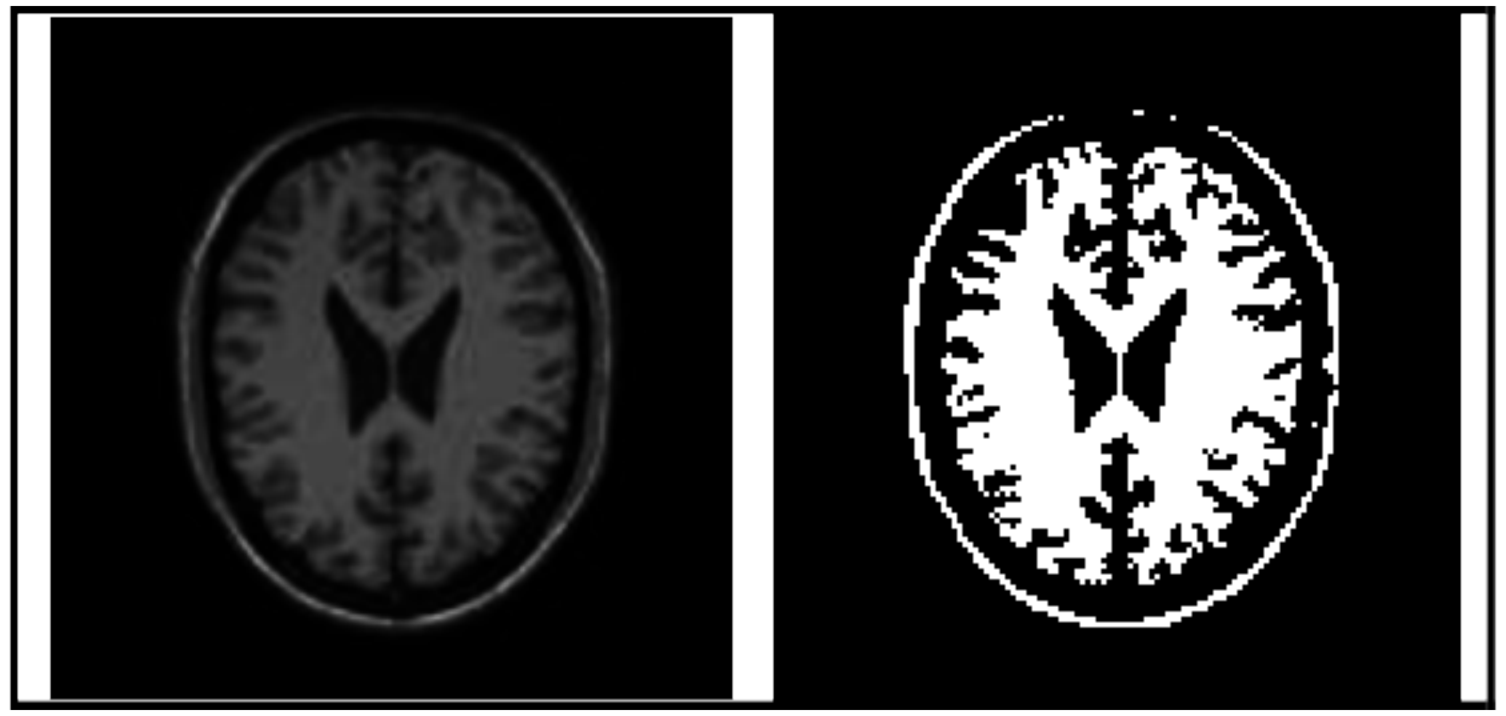
https://github.com/RTLPHD/EmergencyDepartmentForecasting

Methods

Approaches

- Threshold-based data segmentation using specific predictor variables.
- Cluster-based ensemble learning with machine learning models trained for each cluster.

Clustering example for 5 clusters (K-means)

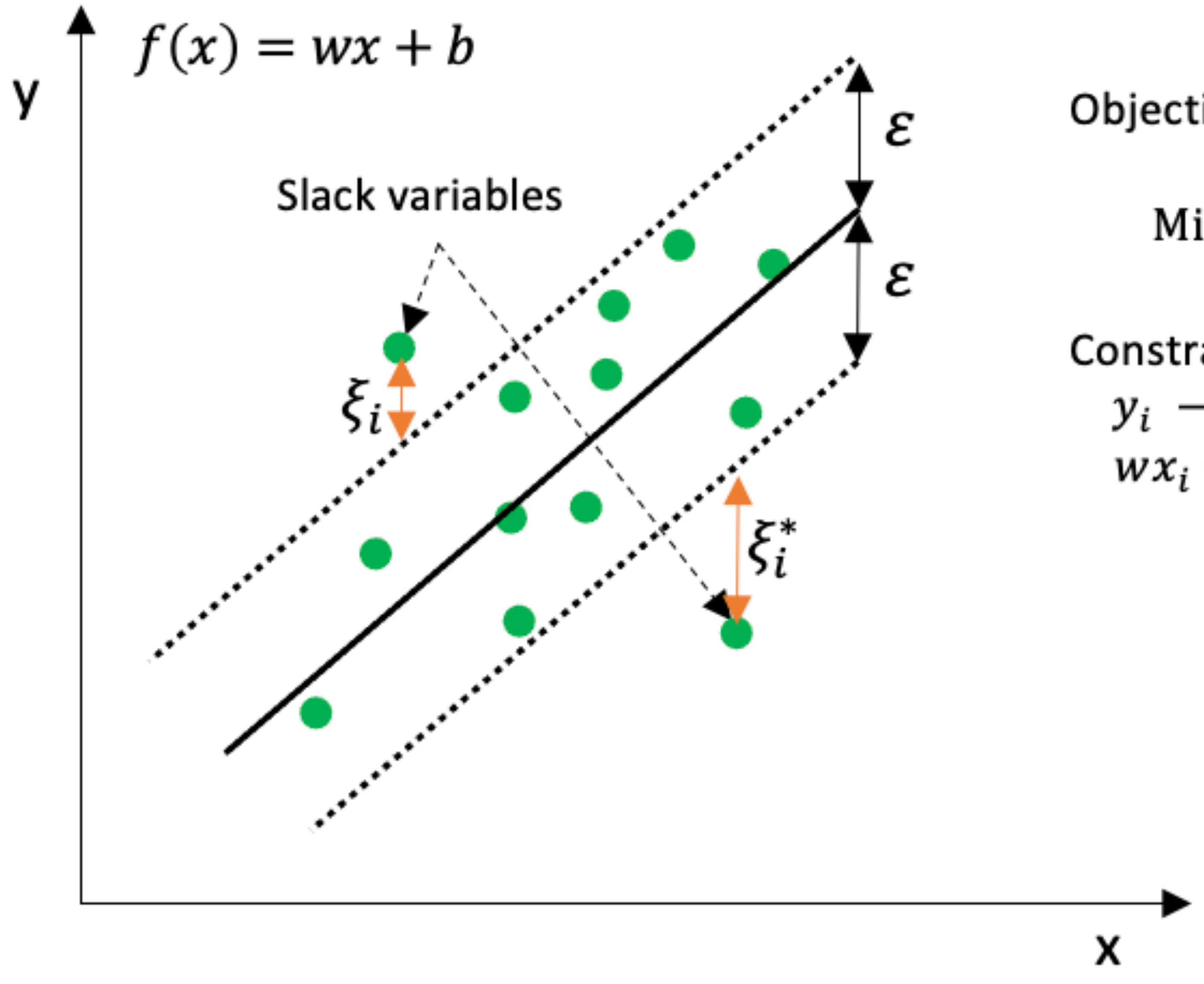


Threshold-based segmentation example on image

Methods

Models

Predicting the number of emergency
department (ED) attendance with
linear regression, metric regression(distance of points with parameters)
and support vector regression (SVR) (for errors and NL)



Objective:

Minimize: $\frac{1}{2}||w||^2 + \mathrm{C}\sum_{i=1}^{l}(\xi$

Constraints:

$$y_i - wx_i - b \le \varepsilon + \xi_i$$
$$wx_i + b - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i^*, \xi_i^* \ge 0$$

Univariate linear SVR (allowing for errors)

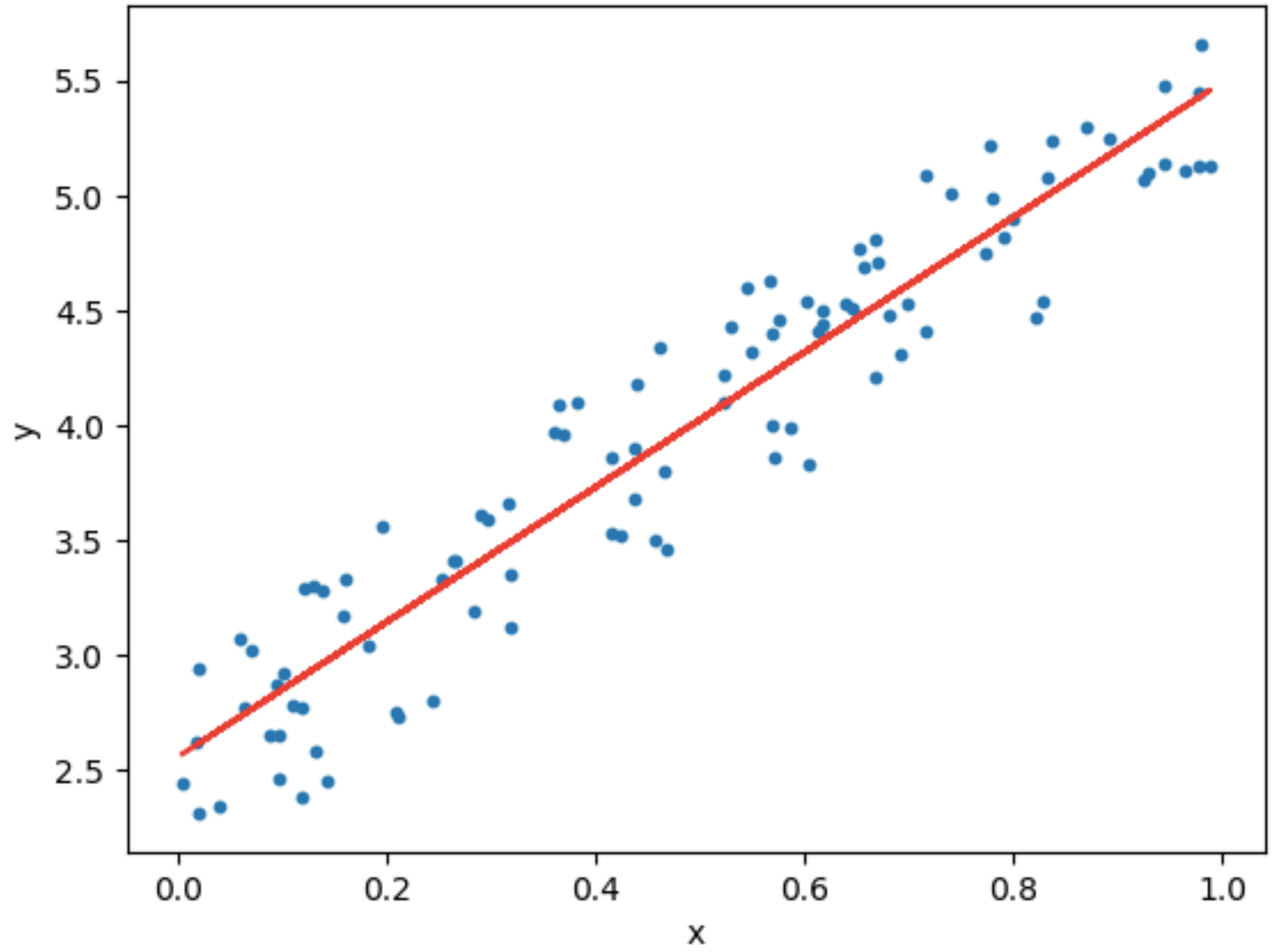
$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,
$\hat{y}$ – *predicted value of y*
$\bar{y}$ – *mean value of y*



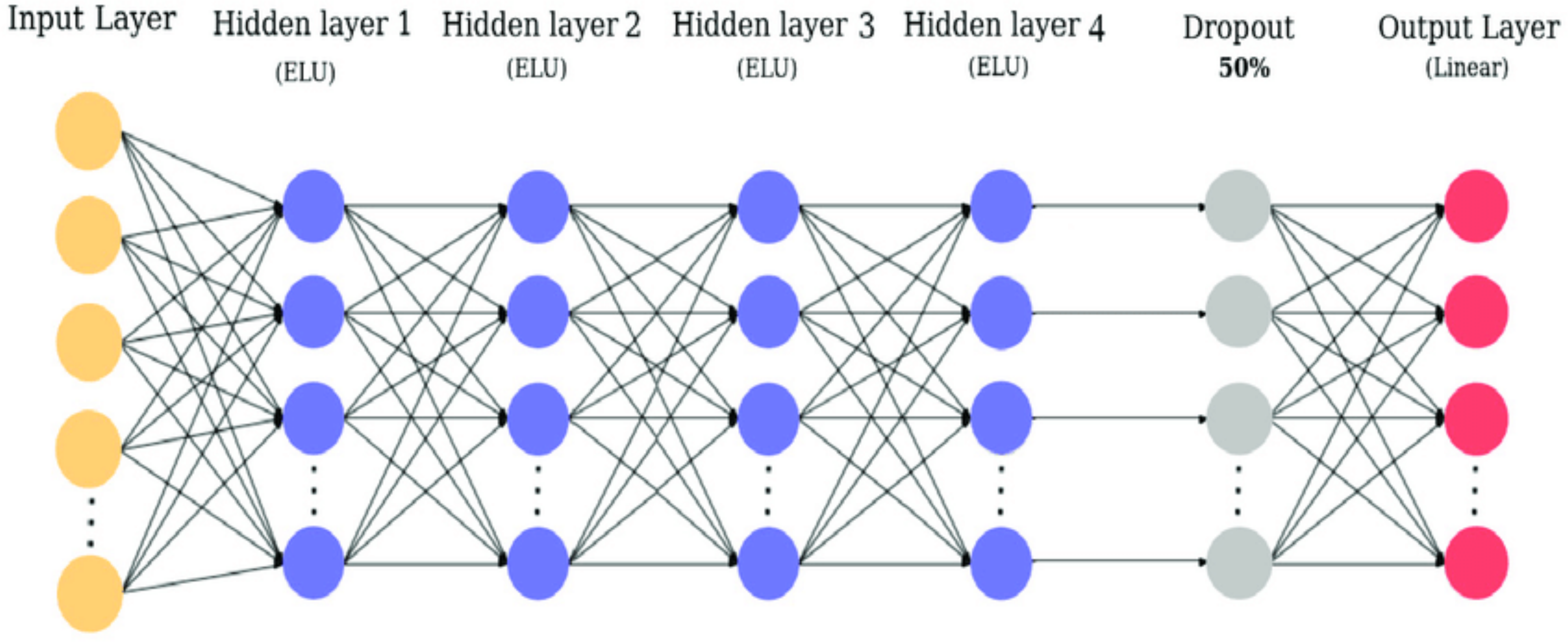
Linear regression example

Methods

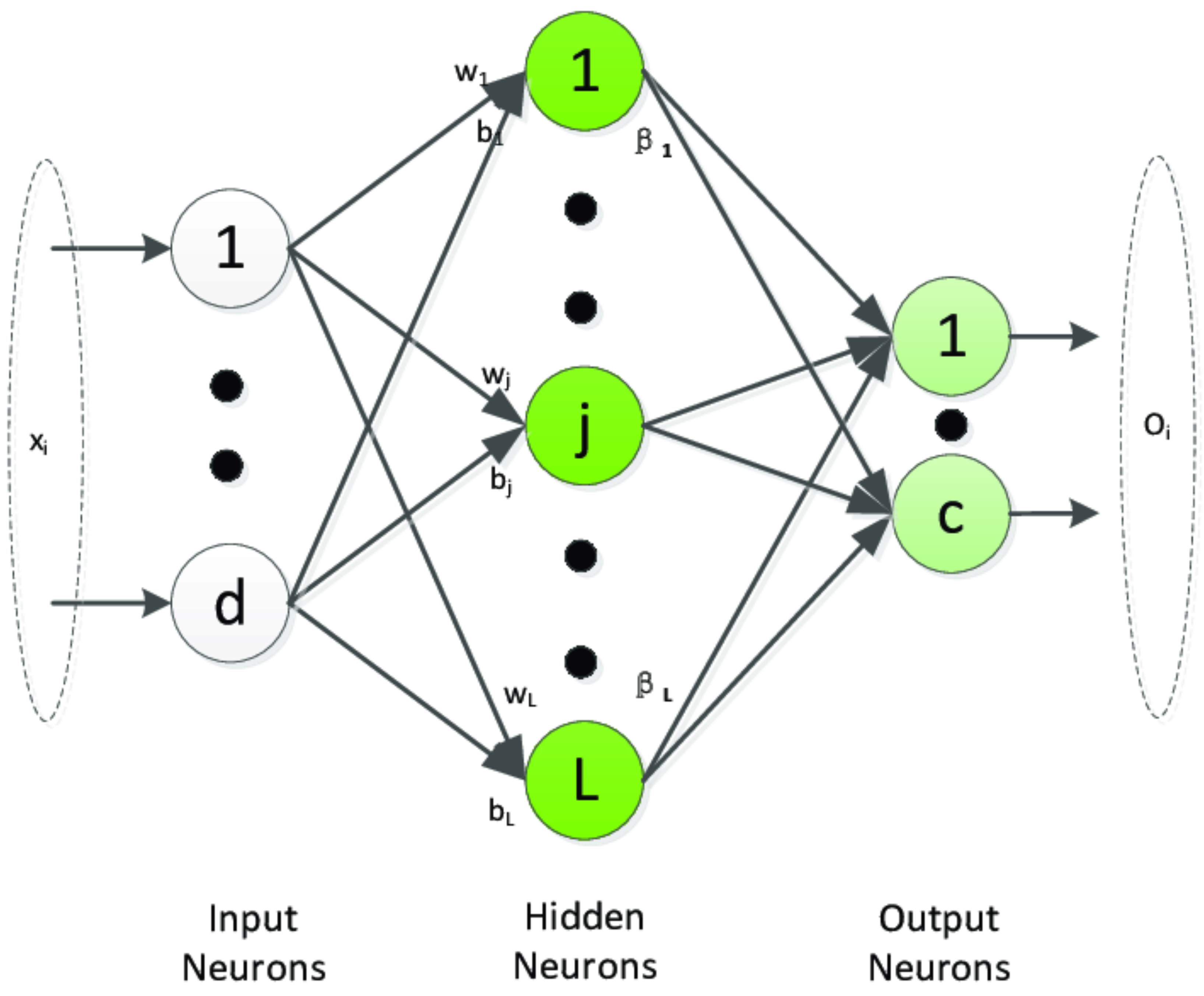
Models

To improve the accuracy of long-term forecasts based on dynamically changing data in ED
Used in the experiment

- Extreme Learning Machine (ELM) and
- Fully Connected Deep Neural Network



FCDNN architecture



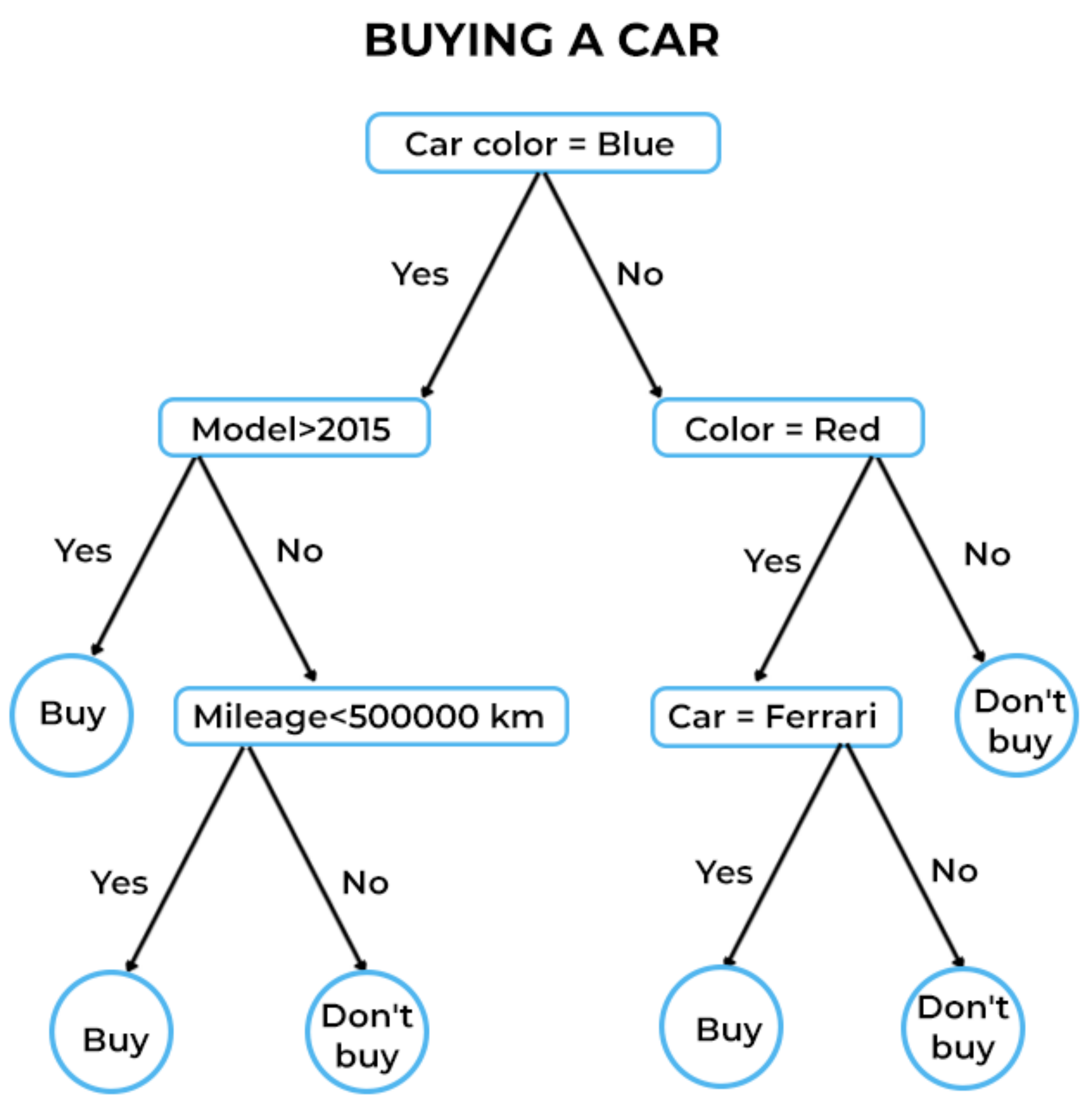
ELM architecture

Methods

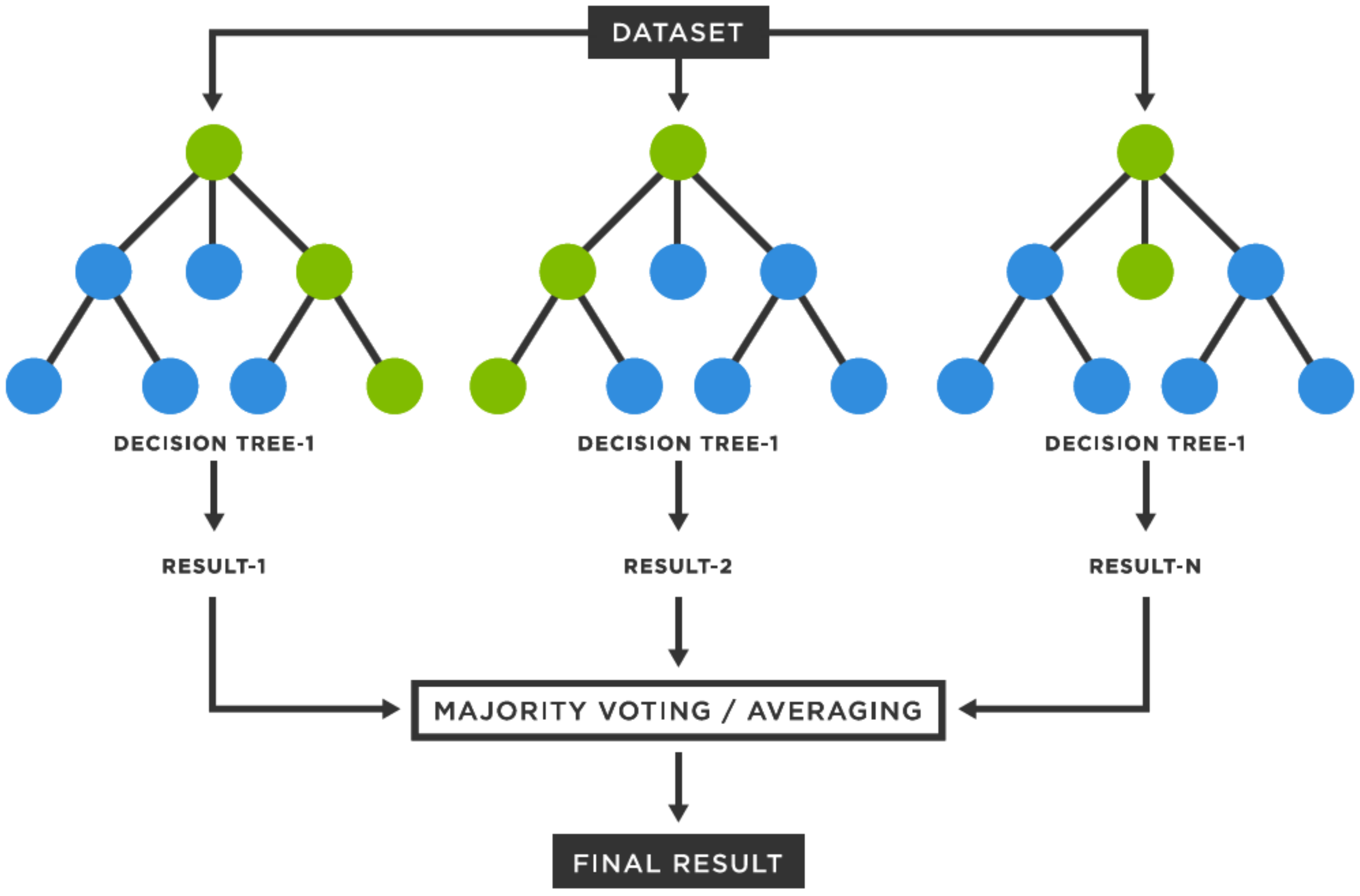
Models

Methods to stop overfitting

- Random forests
- Decision tree



Decision tree example



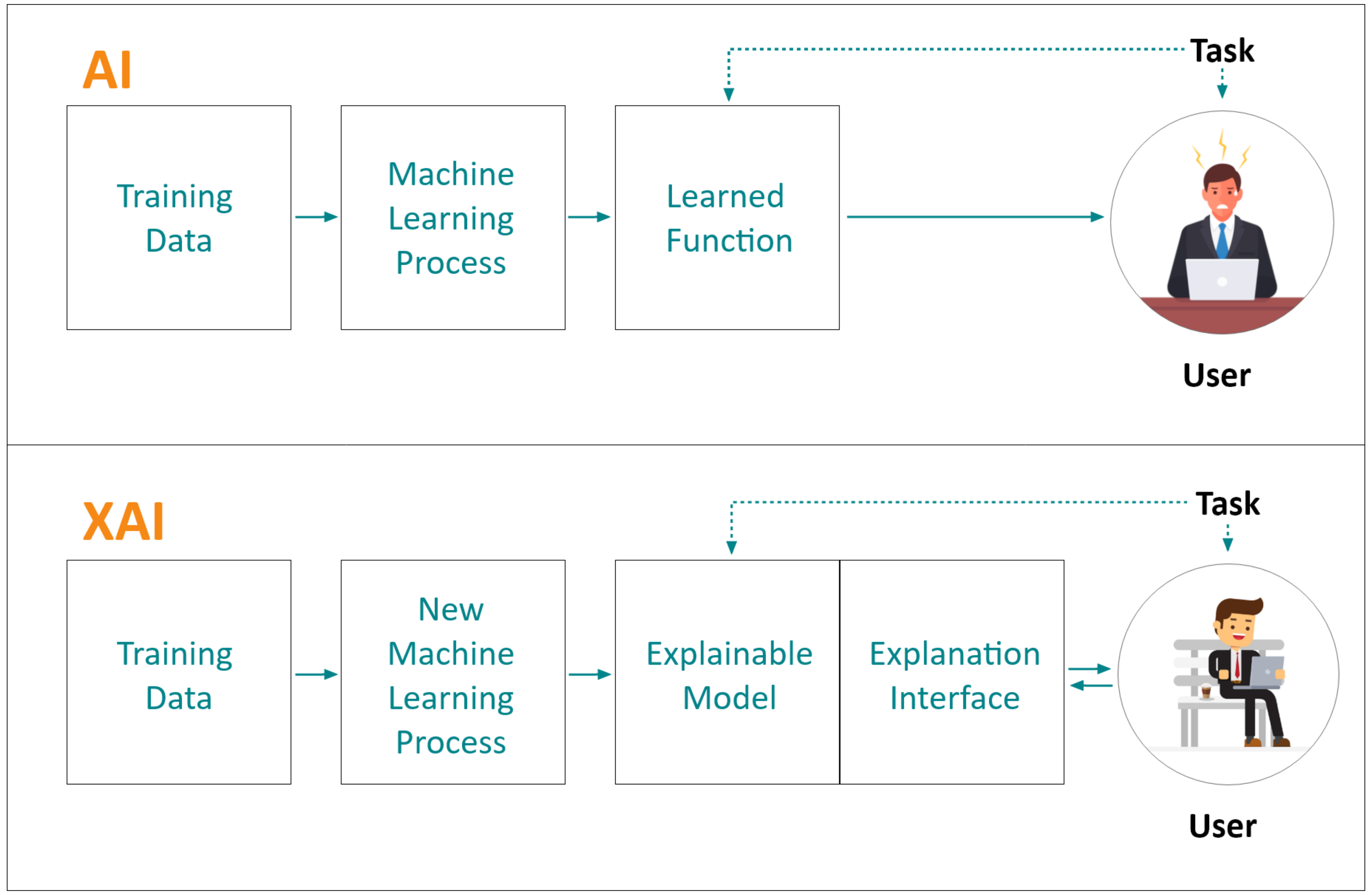
Random forest architecture

Methods

XAI

Explainable AI (XAI) - ensures transparency in predictions.

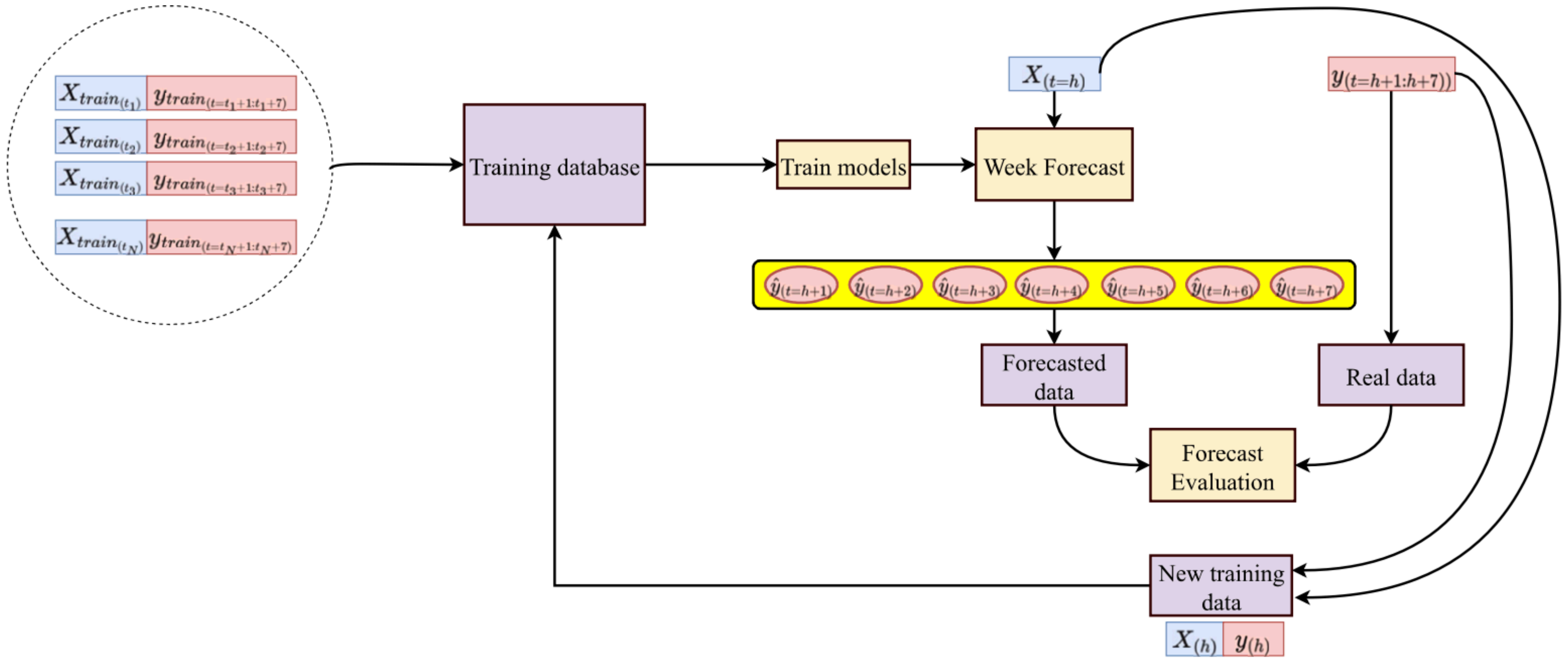
- Highlights key factors driving results (e.g., weather, day of the week).
- Builds trust among ED workers.
- Helps decision-makers understand and confidently use the model.



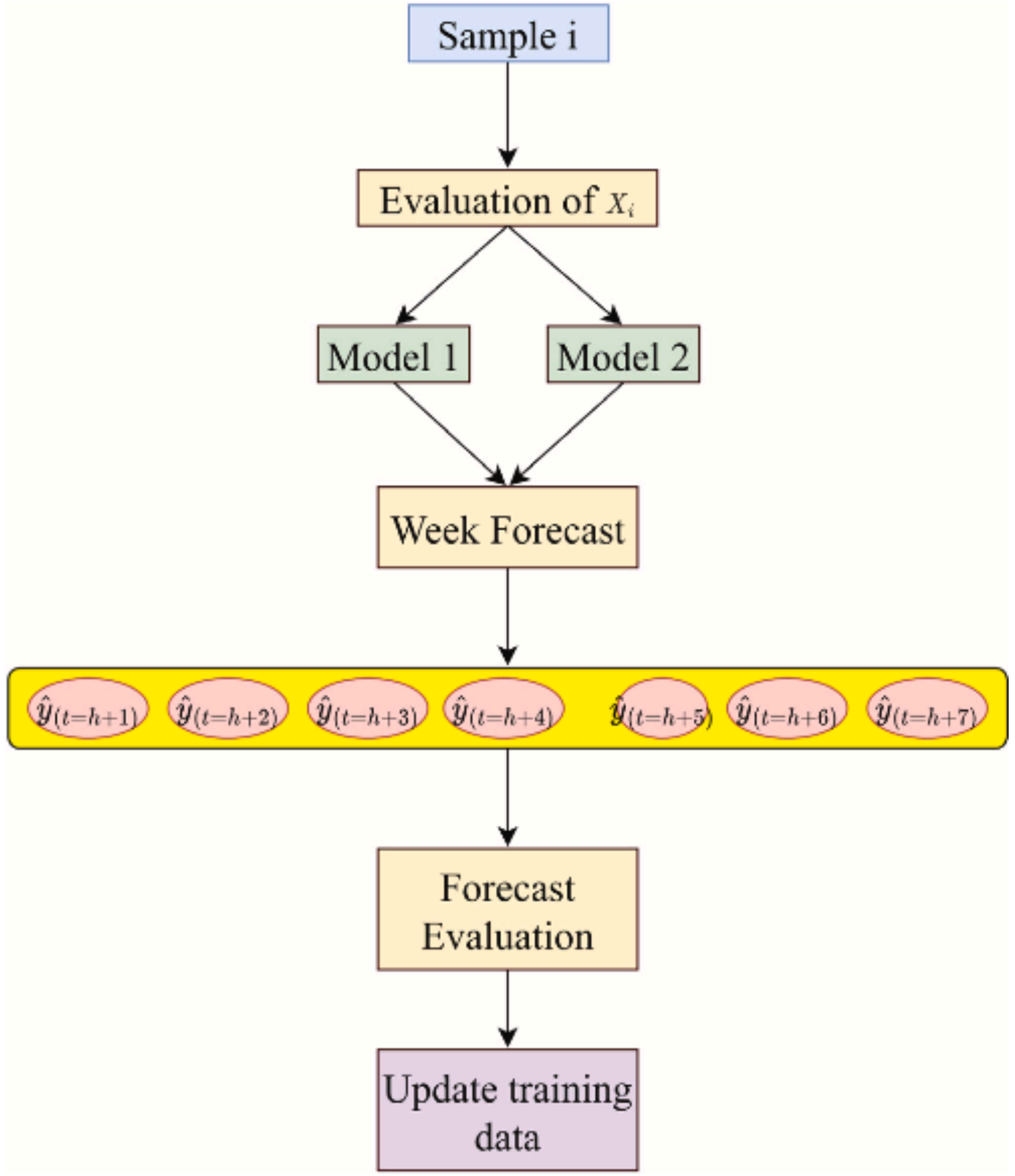
Model

Changing and emergence of new data in real time require model adaptability

So solution is - **continuous training**



Continuous training model



Prediction procedure after the training data segmentation has been performed.

Experiment

Set-up

A. Database construction

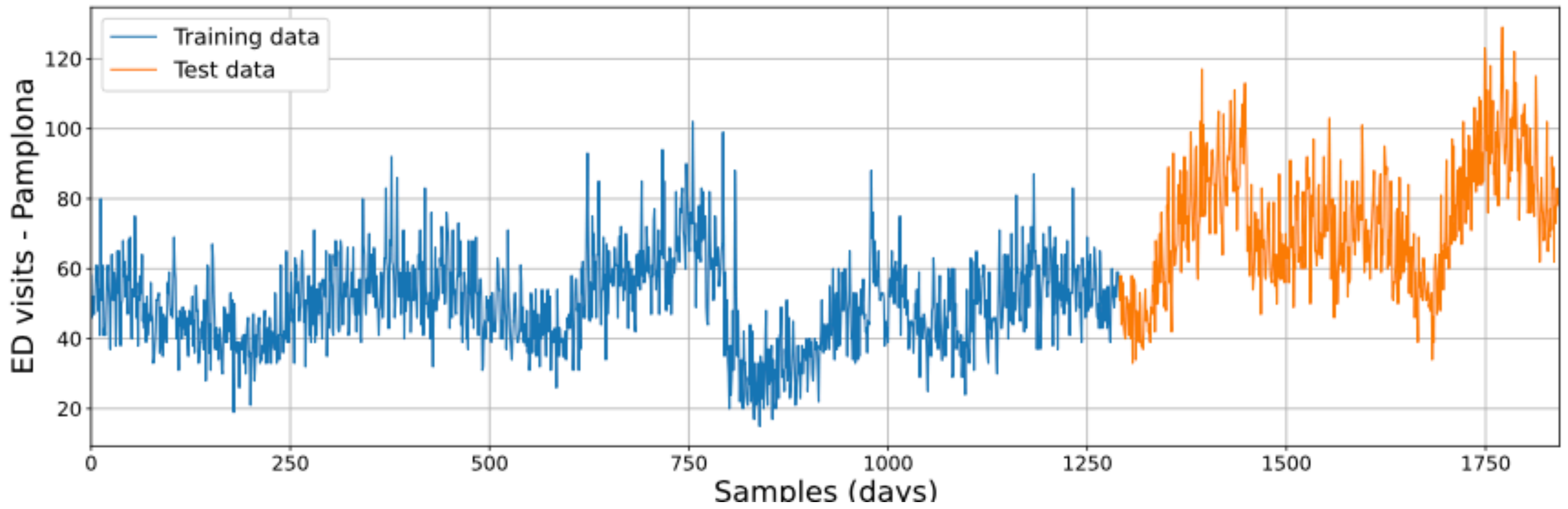
B. Training test split

C. Parameters

D. Metric regression

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

Metric regression to the two proposed problems MAE



Pamplona time series

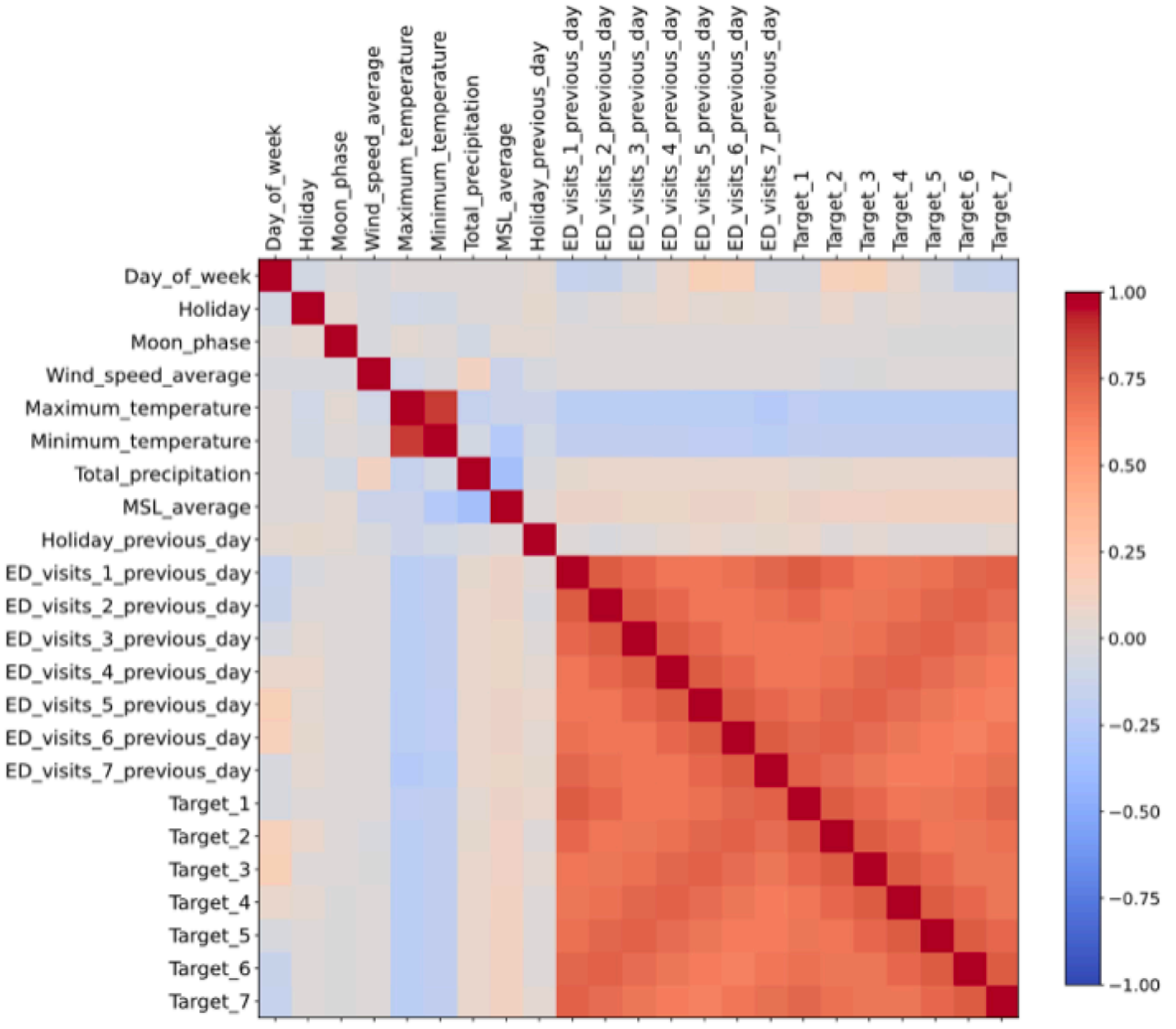
Day_of_week
Holiday
Moon_phase
Maximum_temperature
Average_temperature
Average_Wind_speed
Maximum_Wind_speed
MSL_average
Total_precipitation
Holiday_previous_day
ED_visits_1_previous_day
ED_visits_2_previous_day
ED_visits_3_previous_day
ED_visits_4_previous_day
ED_visits_5_previous_day
ED_visits_6_previous_day
ED_visits_7_previous_day
Target_1
Target_2
Target_3
Target_4
Target_5
Target_6
Target_7

Special parameters for regression

Experiment

- Real ED visitors datasets from Pamplona and Madrid branch
- Including meteorological, calendar and autoregressive principles.
- Testing and training data’s splitting to 70% and 30%
- Clustering to improve accuracy
- 6 ML models including CT

h+6, h-3 till h-1



Correlation coefficient among the variables belong to Pamplona DS

Results and discussion

- Continuous training improved prediction accuracy by 8-19% for Pamplona and 3-5% for Madrid.
- Data segmentation methods have shown better results in predictions compared to baseline models.
- Data clustering provided additional improvements in certain scenarios.
- Using SVR and metric-regression to improve forecast accuracy

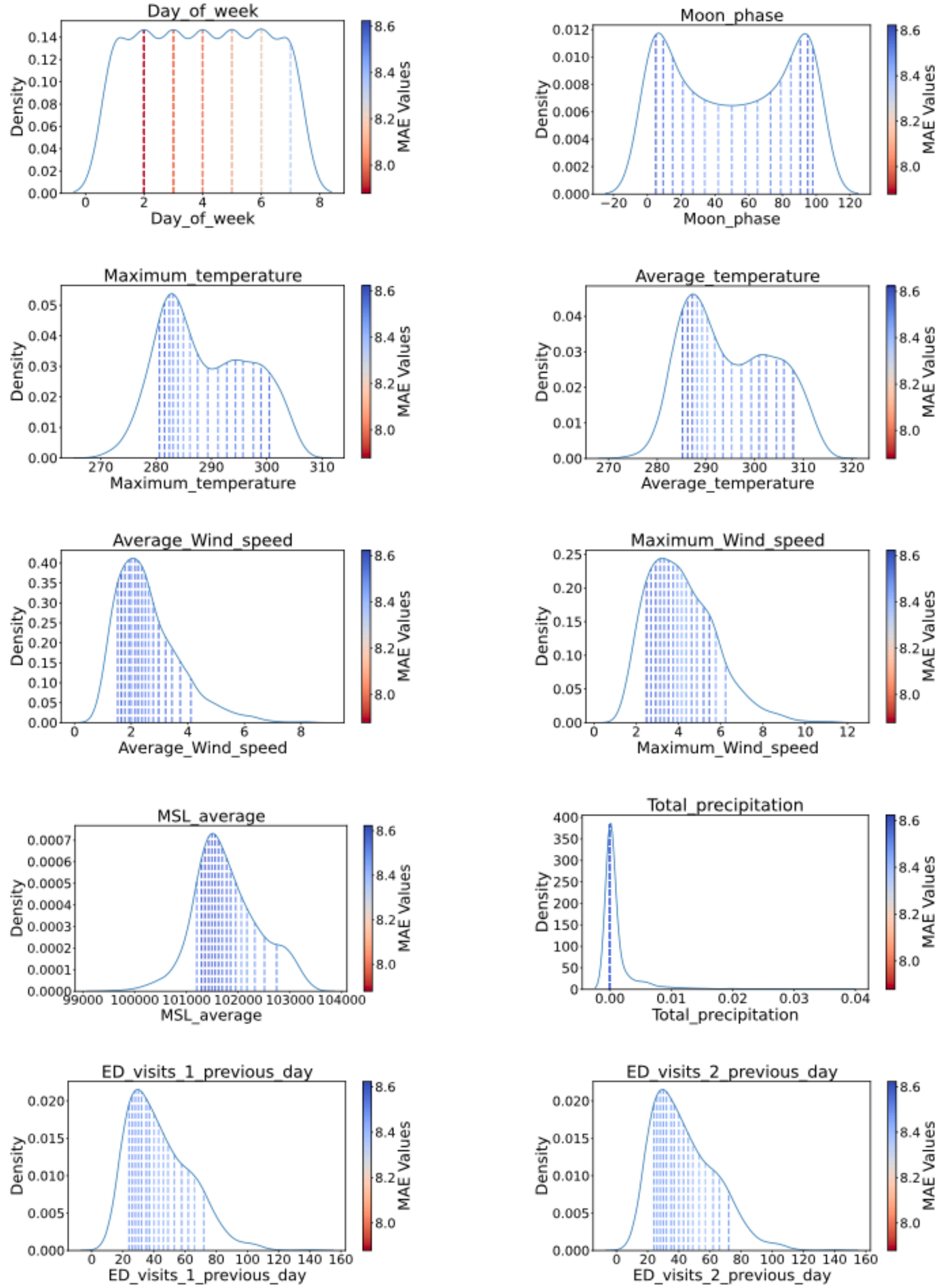
Results (MAE) for Madrid database after applying the continuous training approach.

	+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%

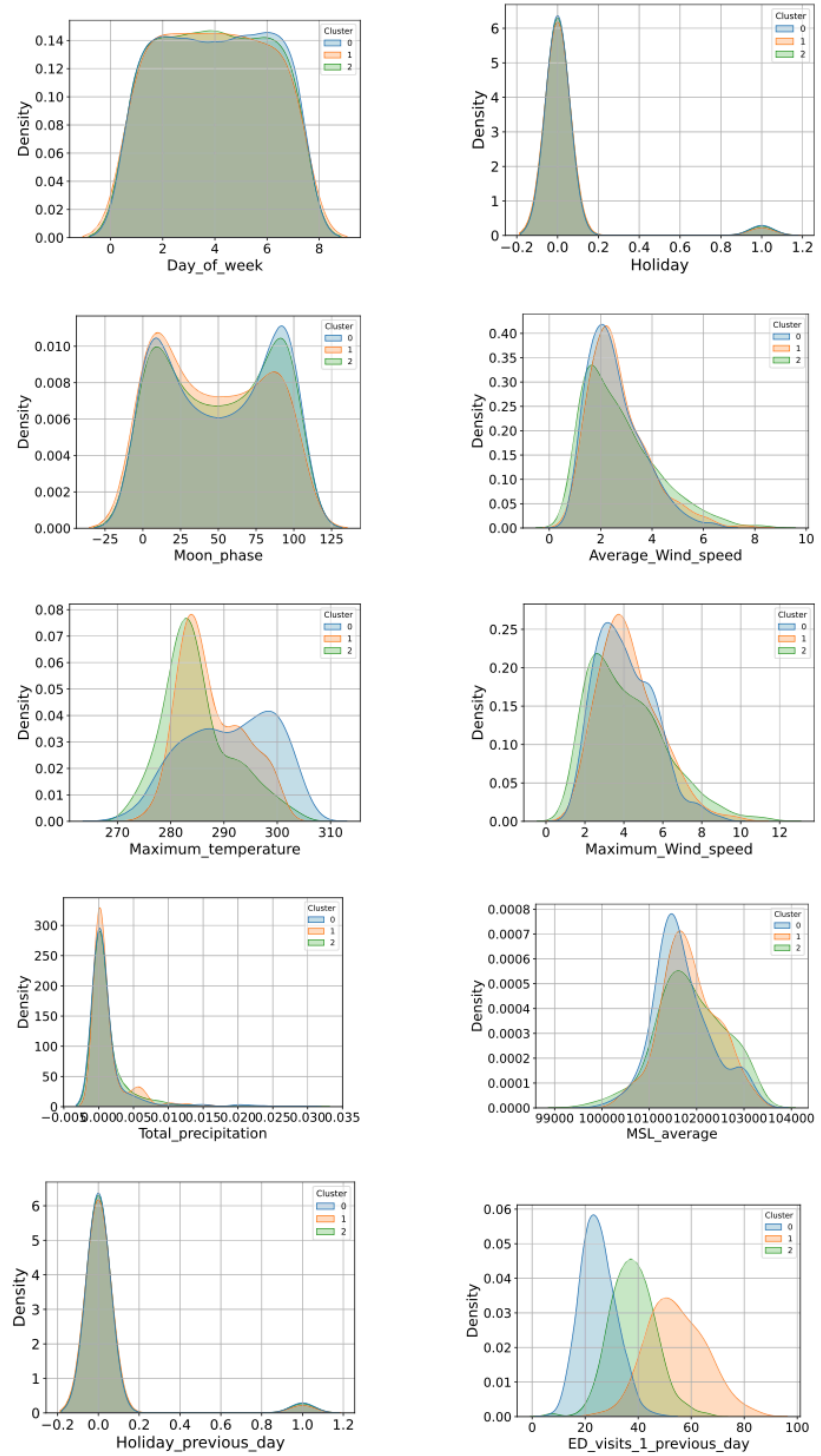
Results (MAE) for Madrid database applying the threshold-based segmentation approach.

	+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%

Results and discussion

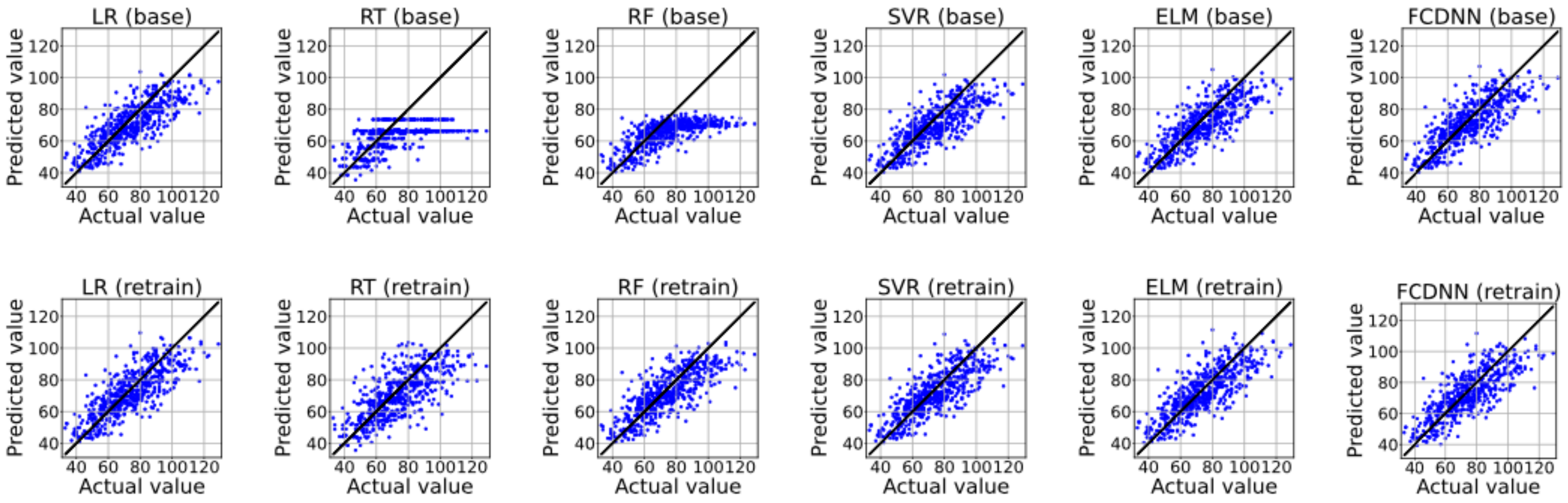


Threshold-based segmentation



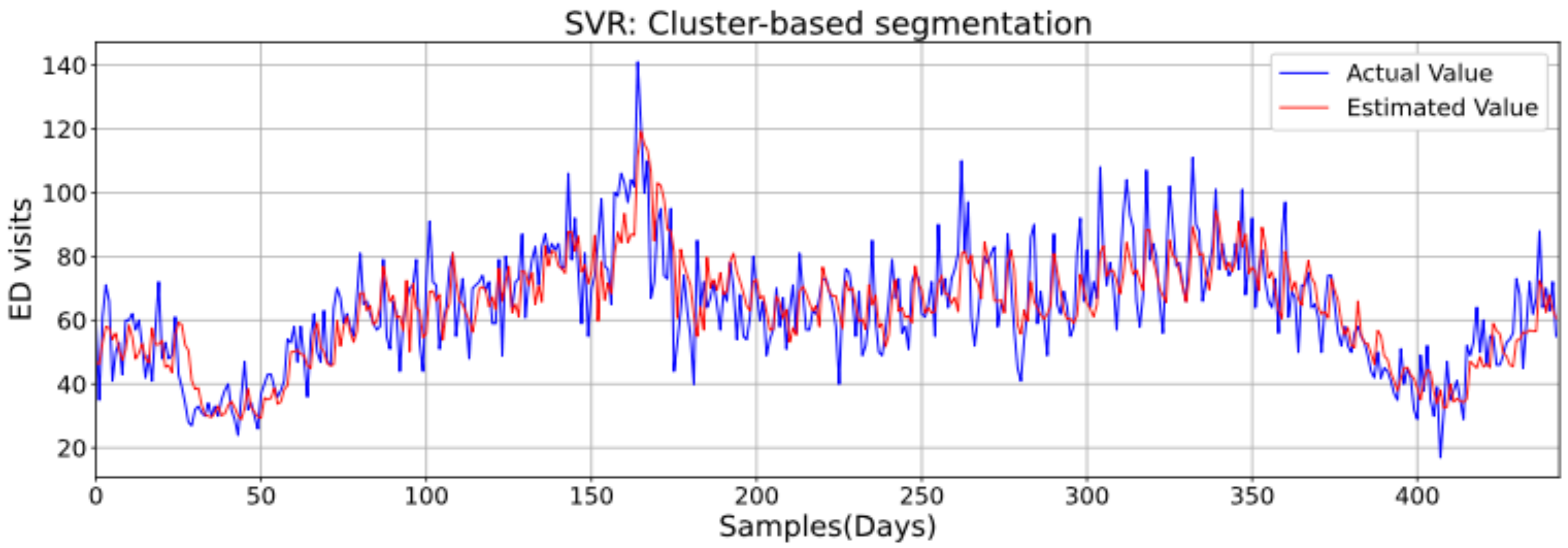
Clustering (3) Madrid

Results and discussion

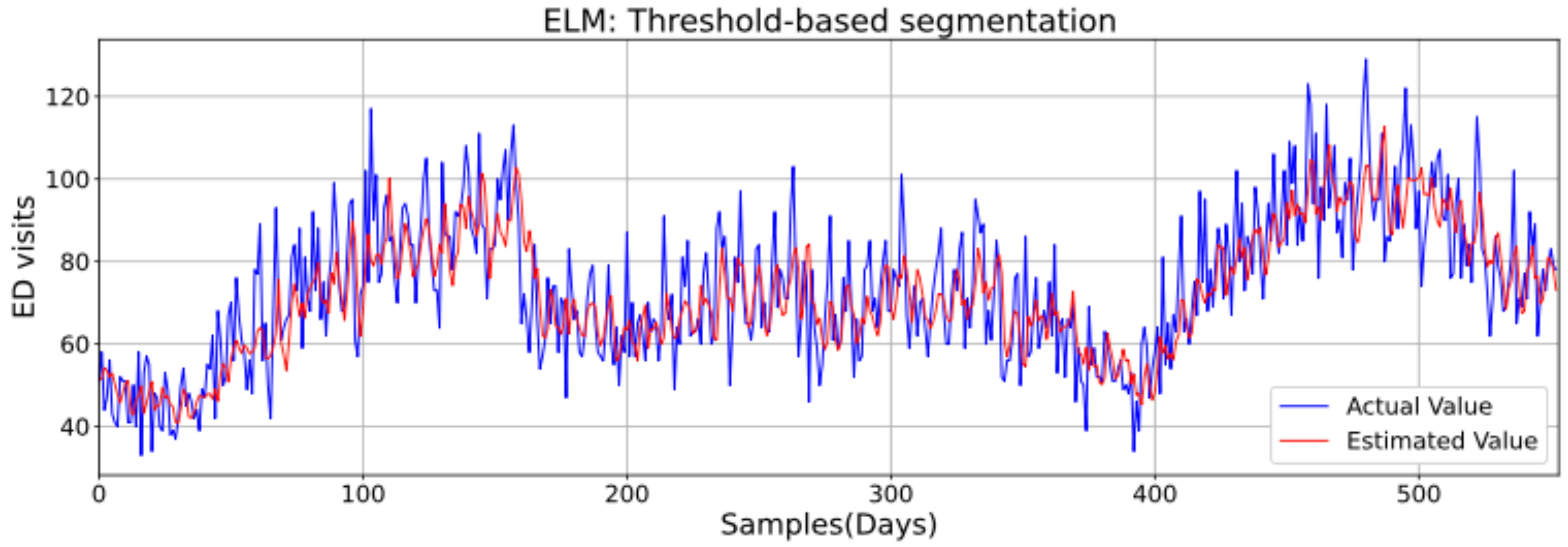


Base models vs continuous training

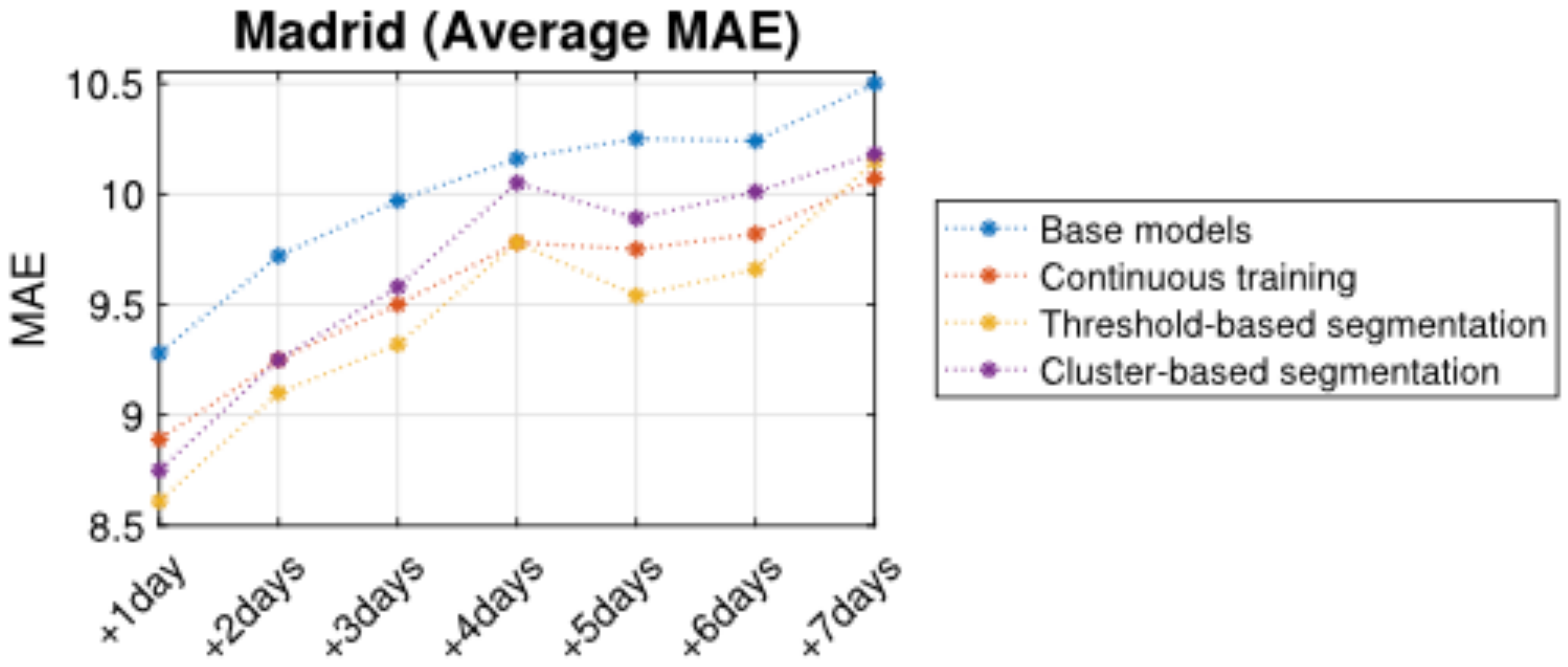
Results and discussion



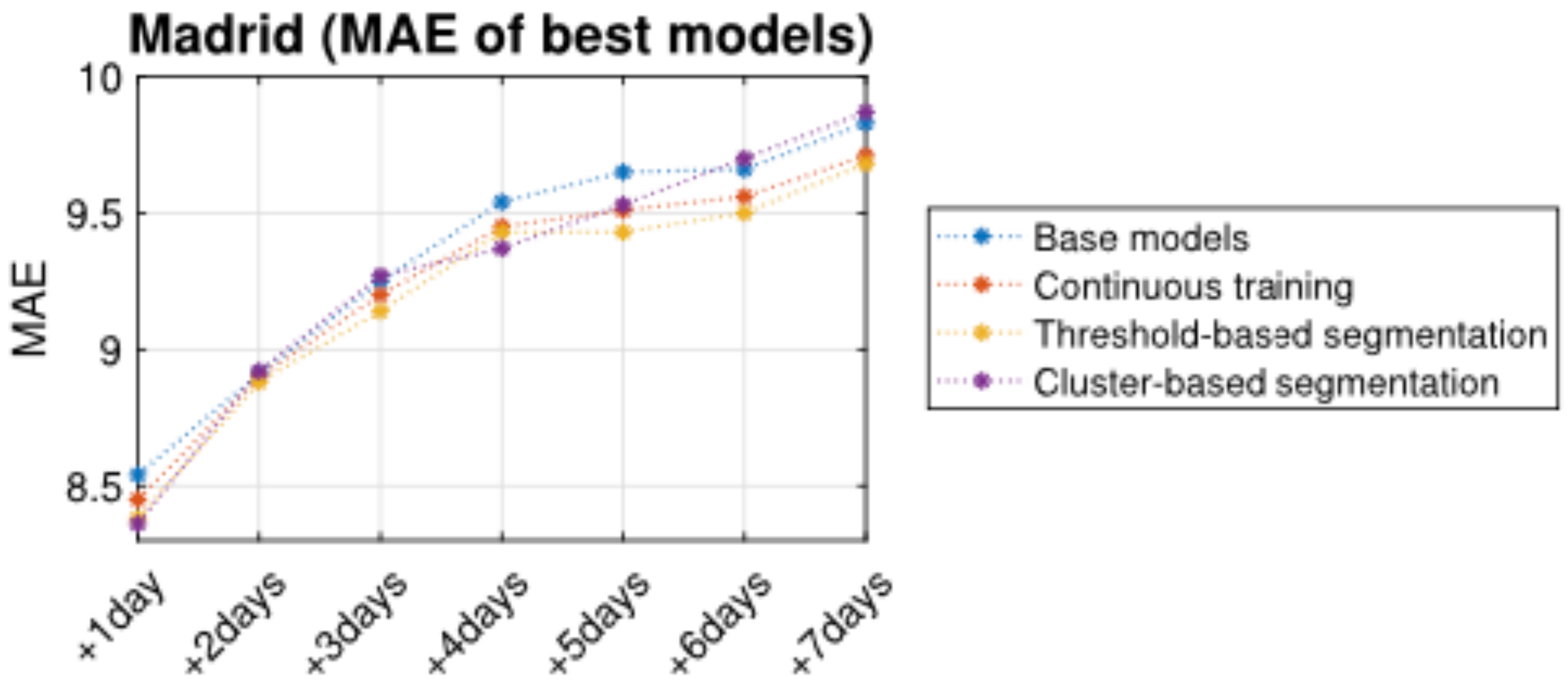
Final results (real vs estimated) Madrid (SVR)



Final results (real vs estimated) Madrid (ELM)



Averaged performance evaluation of the six ML models, assessed in terms of MAE



Comparison of MAE results for the top-performing models acquired through each methodology for each time horizon

Conclusion

- Improved prediction precision by 8-19% for Pamplona and 3-5% for Madrid
- Threshold-based segmentation enhanced model performance by up to 10%.
- Cluster-based models provided reliable accuracy with data-specific clustering.
- Continuous training improved real-time forecasting and adaptability to new data.
- Increased efficiency in hospital resource management and scheduling.

Thank you for listening!

Salimli A.

salimli.am@edu.spbstu.ru