Journal of Sciences, Islamic Republic of Iran 35(1): 63 - 69 (2024) http://jsciences.ut.ac.ir
University of Tehran, ISSN 1016-1104

Prediction of Anoxic Tonic Seizures due to Asthma in Children Using Machine Learning Methods

K. Motarjem[1] and M. Moghimbeygi[2*]

[1] Department of Statistics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Islamic Republic of Iran
[2] Department of Mathematics, Faculty of Mathematics and Computer Science, Kharazmi University, Tehran, Islamic Republic of Iran

Received: 22 February 2024 / Revised: 26 june 2024 / Accepted: 27 July 2024

Abstract

The objective of this study is to investigate the factors influencing asthma attacks in children under six years old using machine learning (ML) methods. There are many statistical methods for data classification that can be used to classify medical data. But using the data itself as well as a set of different methods in machine learning can provide vast and more comparable results. Hence, this study applied ML approaches to predict asthma and second anoxic tonic seizures due to asthma (ATSA) based on variables such as first ATSA, age, region of residence, parent smoking status, and parents' asthma history. The results revealed that children's age and place of residence significantly affected the duration of asthma attacks, with children living in certain areas of Tehran experiencing shorter intervals between attacks due to high air pollution. Machine learning techniques proved useful in predicting ATSA based on age, gender, living region, parents' smoking status, and asthma history, with the AdaBoost method highlighting the importance of the child's age and living area in predicting ATSA.

Keywords: Asthma; Childhood; Prediction Model; Machine Learning.

Introduction

Asthma is a chronic lung disease caused by inflammation in the airways (1). It is the most common chronic disease, affecting 7.1 million (9.6%) of American children. Statistics show that in the United States alone in 2008, children with asthma accounted for 9.3 billion, or 8% of total direct health care costs (2). Symptoms begin in about 80% of children with asthma before the age of six. However, only about 1/3 of children who have at least one episode of asthma symptoms by age three (3) develop asthma at age six or more (4), that is, about 97% of children who have asthma under the age of 3 will not have asthma at the age of six years (5).

Presenting and developing a model to predict whether a child will develop asthma in the future is one of the most critical issues and interests of researchers in children's asthma studies. Such a model can offer several advantages. The most important point is that timely diagnosis and treatment of asthma can prevent serious complications of asthma (6). These allow children to enjoy long-term benefits such as fewer respiratory

* Corresponding Author: Tel: +989126659927; Email: m.moghimbeygi@yahoo.com

symptoms and reduced doses of asthma control drugs (7), even if the treatment is not complete, and as a result, they have fewer drug side effects. Another advantage of obtaining a model for children's asthma is that the diagnosis of this disease is subjective for children under five years of age by doctors (8), and there is no definitive test or genetic test to definitively diagnose it in children. Developing a model can be of great help in this regard. Finally, receiving the model can directly affect the quality and lifestyle of children, because by knowing the severity of the disease, appropriate recommendations can be made to children and their parents or caregivers, from nutrition to exposure to pollution. Identifying and classifying children who are potentially more exposed to asthma can also be one of the advantages of children's asthma prediction modeling. Different approaches may be considered for this modeling, one of the most important of which is building a model based on machine learning methods.

Among the modern methods of statistical analysis, the use of machine learning (ML) has been increasingly used and attracted the attention of researchers to analyze healthcare data and help understand the heterogeneity of asthma and predict its progression. In studies of pediatric diseases, machine-learning approaches have significantly improved the discovery of asthma phenotypes in clinical research. In addition, using machine learning, several accurate models have been introduced to predict asthma (9) and the duration of hospitalization of patients in the emergency room (10). Machine learning is the study of computer algorithms that automatically improve through experience. In this sense, machine learning is an umbrella that includes all computational methods designed to learn through experience (available data) to improve performance and make accurate predictions.Unsupervised machine learning examines and identifies data patterns without predefined outcomes, and supervised machine learning involves learning a rule to predict an outcome based on input-output samples.

In the case of unsupervised machine learning, data-driven approaches using clustering methods can help characterize heterogeneous diseases' diseases among distinct patients. By revealing the underlying structure of the data, cluster analysis can identify a set of samples in the collection of different clusters. For example, the k-means algorithm is one of the most popular iterative descending clustering methods, which aims to minimize the sum of variance within clusters and maximize the separation between clusters, thereby identifying distinct groups in the population. In contrast, aggregate hierarchical clustering algorithms follow a tree structure where the initial nodes represent the instances to be clustered, and the root node represents a supercluster containing all instances. In the case of supervised machine learning, various classifiers have been implemented using regression or classification methods.

The most used methods are linear regression for quantitative response variables and logistic regression for categorical responses. However, in the era of big data , the potential of using machine learning methods has increased significantly, and more advanced models are being developed. In many studies, different characteristics of asthma have been identified to distinguish clusters. The characteristics of the age of onset, allergic sensitivity, severity, and exacerbation in the previous year in 613 children with asthmaled to the identification of 5 different asthma phenotypes, which were reviewed in (11). Clinical and functional characteristics were combined with gene expression profiles of 351 children with asthma to derive five distinct phenotypes of childhood asthma, including lung function, frequency of symptoms, healthcare utilization, percentage of eosinophil, and neutrophils in peripheral blood and serum IgE by (12) in Taiwan. In addition, five different latent classes were identified in demographic characteristics, asthma control, sensitivity, type 2 inflammatory markers, and lung function (13). Although these studies provide valuable information to childhood asthma researchers, it should be noted that the results of these “data-driven” approaches need more validation to expand knowledge about asthma phenotypes in children. Therefore, many features are effective in the occurrence of asthma, and these features can be suitable tools in machine learning methods for the necessary predictions.

Considering that most of the existing models for predicting the development of asthma in children are based on the logistic regression statistical method, other existing models are based on the risk score or a combination of risk factors. As with predictive modeling in general, machine learning methods such as support vector machines and random forests often achieve higher predictive accuracy than risk score, a combination of risk factors, and logistic regression. Comparing different machine learning methods to predict the development of asthma in children is an important issue.

In this article, we are going to examine different machine-learning methods for analyzing children's asthma data. Next, in the second part, ten commonly used models in machine learning are implemented on the data related to the asthma of children under six years old in Tehran. Finally, a discussion and conclusion will be presented. In this study, the seizure status of children aged 2 to 6 years with asthma living in Tehran who were referred to one of the clinics in this city has been investigated.

Material and Methods

The data are related to 208 children with asthma in the age range of 16 to 66 months in Tehran. More details about the data are available in (14). The data were collected by monitoring children entered into the study after their first ATSA based on age, gender, the second ATSA, and place of residence at one year. This study aims to classify children based on their second ATSA concerning other feature variables. The analyses of this study were done using the Orange software and based on machine learning methods. The necessary details about the variables are given in Table 1. Since the AdaBoost and Gradient Boosting methods performed better in data analysis, we will discuss these methods in more detail in the next section.

Adaptive Boosting (AdaBoost)

The AdaBoost algorithm was introduced in 1995 by Freund and Shapier, which solved many of the practical problems of previous boosting algorithms (15). The generic strategy in this method is based on combining classifiers with simpler classifiers. The main idea of this algorithm is to select a "weak classifier" and use it to build a better classifier, thereby increasing the performance of the weak classification algorithm. This improvement is done by averaging the output of a set of weak classifiers. The most popular boosting algorithm is AdaBoost, which is called "adaptive". This algorithm is much easier to use and implement than SVMs and often provides better results. There is also great flexibility in choosing weak classifiers. Boosting is a special case of a general class of learning algorithms called ensemble methods that try to create better learning algorithms by combining several simpler algorithms.

Let $\{(x_i, y_i)\}$ for $i = 1,2, \ldots N$ as training data where $x_i \in R^K$ and $x_i \in \{-1,1\}$. Suppose we are given a (potentially large) number of weak classifiers, denote $f_m(x) \in \{-1,1\}$ and 0-1 loss function I, defined as

$$I(f_m(x), y) = \begin{cases} 0 \text{ if } f_m(x_i) = y_i \\ 1 \text{ if } f_m(x_i) \neq y_i. \end{cases}$$

Then, the pseudocode of the AdaBoost algorithm is as follows:

for i from 1 to N, ω_i =1
for m = 1 to M do
Fit weak classifier m to minimize the objective function:

$$\epsilon_m = \frac{\sum_{i=1}^{N} \omega_i^{(m)} I(f_m(x_i) \neq y_i)}{\sum_i \omega_i^{(m)}}$$

Where $I(f_m(x_i) \neq y_i) = 1$ if $I(f_m(x_i) \neq y_i)$ and 0 otherwise

$$\alpha_m = \ln \frac{1 - \varepsilon_m}{\varepsilon_m}$$

for all i do

$$\omega_i^{(m+1)} = \omega_i^{(m)} e^{\alpha_m I(f_m(x_i) \neq y_i)}$$

end for
end for

After learning, the final classifier is based on a linear combination of the weak classifiers:

$$g(x) = \operatorname{sign}\left(\sum_{m=1}^{m} \alpha_m f_m(x)\right)$$

Indeed, AdaBoost is a greedy algorithm that makes up a "strong classifier", i.e., g(x), incrementally, by optimizing the weights and adding one weak classifier at a time.

Gradient Boosting

Another classification method is the use of loss functions and basic learning models in the optimization algorithm. In practice, given some specific loss function $\Psi(y, f)$ and/or a customized base learner $h(x, \theta)$, it can be difficult to obtain a solution for parameter estimation. To counter this, it is proposed to choose a new function $h(x, \theta_t)$ to be the most parallel to the negative gradient $\{g_t(x_i)\}$ for $i = 1, 2, 3, \ldots, N$ along the observed data:

Table 1. The data summary

	Details
Subjects	Anoxic tonic seizures due to asthma
Source	Motarjem et.al (2018)
Year	2017
Monitoring period	One year
Sample size	208
Sex	119 male(1), 89 female(0)
Age	Between 16 to 66 months
Region	22 regional municipality of Tehran
Parent smoke	Smoking (1) or non-smoking parents (0)
Parent asthma	Asthmatic (1) or non-asthmatic parent (0)
Outcome	Second ATSA occurred (1) or not occurred(0)

$$g_t(x) = E_y \left[\frac{\partial\Psi(y,f)}{\partial f(x)} |x\right]_{f(x)=\hat{f}(x)^{t-1}}$$

Instead of looking for the general solution for the boost increment in the function space, one can choose the new function increment to be the most correlated with $-g_t(x)$. This allows replacing a potentially challenging optimization task with a classical least-squares minimization:

$$(\rho_t, \theta_t) = argmin \sum_{i=1}^{N} (-g_t(x) + \rho h(x_i, \theta))^2$$

In summary, we can formulate the complete form of the gradient boosting algorithm as presented in (16). The exact form of the derived algorithm with all the corresponding formulas strongly depends on the design choices of $\Psi(y, f)$ and $h(x, \theta)$. Some common examples of these algorithms can be found in (16).
The Gradient Boost algorithm is as follows:

Inputs:

- Input data $(x, y)\ for\ i = 1,2,3, \dots N$
- number of iterations M
- choice of the loss-function $\Psi(y, f)$
- choice of the base-learner model $h(x, \theta)$

Algorithm
1: initialize $\widehat{f_0}$with a constant
2: for $t = 1$ to M **do**
3: compute the negative gradient$g_t(x)$.
4: fit a new base-learner function $h(x, \theta_t)$
5: find the best gradient descent step-size ρ_t**:**

$$\rho_t = argmin_\rho \sum_{i=1}^{N} \Psi[y_i, \widehat{f_{t-1}}(x_i) + \rho h(x_i, \theta_t)]$$

6: update the function estimate:

$$\widehat{f_t} \to \widehat{f_{t-1}} + \rho h(x, \theta_t)$$

7: end for

This algorithm also maximizes the correlation between the whole network error and the newly created neuron, which makes this comparison more evident.

Results

In this article, ten data mining methods are used for classification. These methods are Neural Network, Naive Bayes, Adaptive Boosting (AdaBoost), Gradient Boosting, Random Forest, Classification Tree (Tree), k-nearest neighbor (kNN), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM) and Logistic Regression. In addition, four categorical variables and a numerical variable were used for classification. The confusion matrix was used to calculate the performances of the classifiers. The confusion matrix is shown in Table 2 ,where:

- TP (True positive): The observation is positive, and is predicted to be positive.
- FN (False Negative): The observation is positive, but is predicted to be negative.
- FP (False Positive): The observation is negative, and is predicted to be negative.
- TN (True Negative): The observation is negative, but is predicted to be positive.

In this table, for example, AdaBoost correctly classifies 206 out of 208, while it misclassified only 2 out of 208. In all models, misclassification is lower than true classification. In general, the models have fewer type I and more type II errors.

To evaluate the data mining methods, we use the following criteria: area under the curve (AUC), classification accuracy (CA), F1-score, precision, and recall which they calculated by:

The result can be seen in Table 3. Based on the precision metric, Gradient Boosting, and AdaBoost methods perform classification better than other methods. Therefore, their accuracy is equal to 0.891 and 0.991, respectively. In addition, among these three models, it can be seen that the best performance is for AdaBoost when we have the model (male).

Table 2. Confusion matrix for classification instances

Model	FP	TN	FN	TP	Correct	Incorrect
Logistic Regression	52	47	36	73	125	83
Naive Bayes	54	45	34	75	129	79
SGD	49	50	27	82	131	77
kNN	70	29	31	78	148	60
Tree	90	9	26	83	173	35
SVM	61	38	17	92	153	55
Neural Network	72	27	20	89	161	47
Random Forest	76	23	9	100	176	32
Gradient Boosting	84	15	8	101	185	23
AdaBoost	99	0	2	107	206	2

Table 3. Performance metrics of the ten data mining models

Model (average)	AUC	CA	F1	Precision	Recall
Logistic Regression	0.678	0.601	0.599	0.6	0.601
Naive Bayes	0.671	0.62	0.618	0.62	0.620
SGD	0.624	0.63	0.623	0.632	0.630
kNN	0.759	0.712	0.712	0.712	0.712
SVM	0.185	0.736	0.732	0.743	0.736
Neural Network	0.883	0.774	0.773	0.775	0.774
Tree	0.932	0.832	0.831	0.842	0.832
Random Forest	0.939	0.846	0.845	0.852	0.846
Gradient Boosting	0.961	0.889	0.889	0.891	0.889
AdaBoost	1	0.99	0.99	0.991	0.990
Model (Female)	AUC	CA	F1	Precision	Recall
Logistic Regression	0.678	0.601	0.556	0.591	0.525
Naive Bayes	0.671	0.62	0.578	0.614	0.545
SGD	0.624	0.63	0.56	0.645	0.495
kNN	0.759	0.712	0.7	0.693	0.707
Tree	0.932	0.832	0.837	0.776	0.909
SVM	0.185	0.736	0.689	0.782	0.616
Neural Network	0.883	0.774	0.754	0.783	0.727
Random Forest	0.939	0.846	0.826	0.894	0.768
Gradient Boosting	0.961	0.889	0.88	0.913	0.848
AdaBoost	1	0.99	0.99	0.98	1.000
Model(Male)	AUC	CA	F1	Precision	Recall
Logistic Regression	0.678	0.601	0.638	0.608	0.670
SGD	0.624	0.63	0.68	0.621	0.752
Naive Bayes	0.671	0.62	0.655	0.625	0.688
SVM	0.185	0.736	0.77	0.708	0.844
kNN	0.759	0.712	0.722	0.729	0.716
Neural Network	0.883	0.774	0.791	0.767	0.817
Random Forest	0.939	0.84	6 0.86	2 0.81	3 0.917
Gradient Boosting	0.961	0.889	0.898	0.871	0.927
Tree	0.932	0.832	0.826	0.902	0.761
AdaBoost	1	0.99	0.991	1	0.982

When considering other classification criteria such as AUC, CA, F1, and Recall, Gradient Boosting and AdaBoost techniques demonstrate superior performance compared to other methods. This is significant given that logistic regression, a statistical method, has a significantly lower level of accuracy. Therefore, the most important application of machine learning methods can be seen here. It can be seen from Figure 1 that the main predictor of ATSA is age in precision. It is followed by region. The effect of each variable can be seen separately in Figure 2. As seen, age and gender (male) have the greatest influence on the response variable. In addition, based on Figure 3, it can be seen that the probability of the second ATSA for instance child living in Region 3 will decrease by 60% by changing to Region 1. This finding highlights the importance of modifying the patient's living environment to reduce the likelihood of a second ATSA. As a result, this study also succeeded in lowering the risk of recurrent ATSA for each individual through relocation. Additionally, comparing the time intervals between two asthma attacks in children and the air pollution map of Tehran shows that children residing in areas 11, 3, 9, 12, and 6 have shorter intervals between asthma attacks due to the high level of air pollution in those areas. The relationship between children's asthma attacks and air pollution by the AQI (The AQI is the index for reporting air quality) index is clearly shown in Figure 4.

Conclusion

Based on the data collected in Tehran, Machine learning techniques were useful in predicting ATSA based on age, gender, living region, parents' smoking status, and asthma background. These techniques included ten different methods; the AdaBoost and Gradient Boosting methods performed better based on the average over-classes model. According to the AdaBoost method, the child's age plays an important role in ATSA. In addition, the living area can affect the probability of ATSA, for example, changing the living

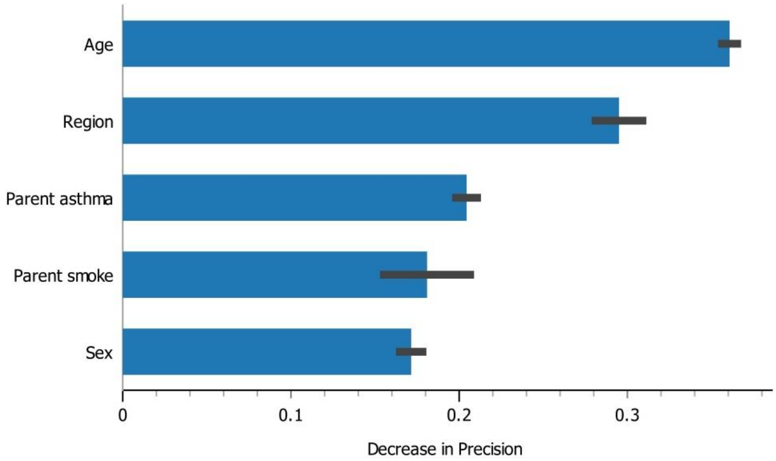

Figure 1. Impact of each feature on precision measure

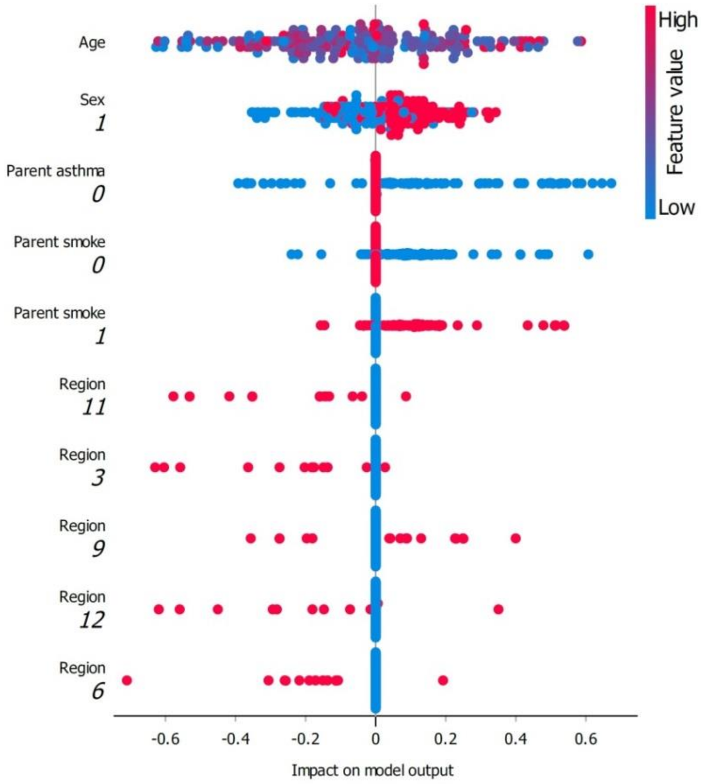

Figure 2. The ranking of the impact of the variables obtained using the AdaBoost Model

region from 3 to 1 decreases the probability of the second ATSA by about 60%. However, differences in data characteristics, variables, and the type of model used can affect the accuracy of the results. Data mining models can be used to design decision support systems that can help reduce ATSA and improve asthma in children. Since machine learning methods are based on the data itself, therefore, a large number of samples can have a positive effect on the performance of the models. As a result, it is strongly recommended to use the data available in other medical centers.

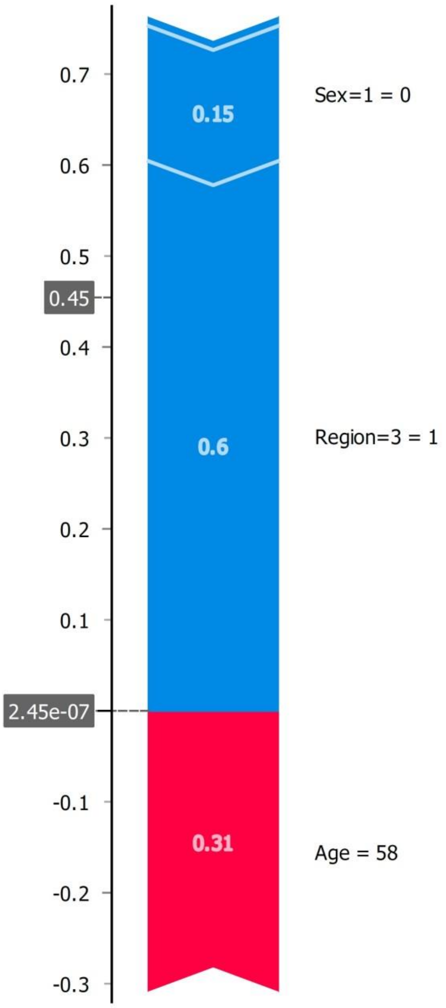

Figure 3. The Probability of second ATSA for an instance child as predicted by the AdaBoost Model

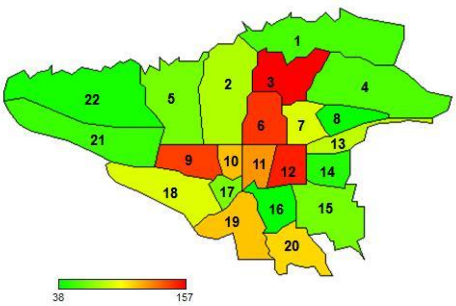

Figure 4. Air pollution map of Tehran city based on one-year average Air Quality Index (AQI) https://airnow.tehran.ir/

References

1. Malveaux FJ. The state of childhood asthma: introduction. Pediatrics. 2009;123 Suppl 3:S129–30
2. Roemer M. Health care expenditures for the five most common children’s conditions. 2008: estimates for U.S. civilian non institutionalized children, ages 0–17. MEPS statistical brief 349. Rockville, MD: AHRQ; 2011
3. Yunginger JW, Reed CE, O’Connell EJ, Melton 3rd LJ, O’Fallon WM, Silverstein MD. A community-based study of the epidemiology of asthma. Incidence rates, 1964–1983. Am Rev Respir Dis. 1992;146(4):888–94
4. Hafkamp-de Groen E, Lingsma HF, Caudri D, Levie D, Wijga A, Koppelman GH, et al. Predicting asthma in preschool children with asthma-like symptoms: validating and updating the PIAMA risk score. J Allergy Clin Immunol. 2013;132(6):1303–10
5. Martinez FD. Development of wheezing disorders and asthma in preschool children. Pediatrics. 2002;109(2 Suppl):362–7
6. Frey U, von Mutius E. The challenge of managing wheezing in infants. N Engl J Med. 2009;360(20):2130–3
7. Haahtela T, Tamminen K, Kava T, Malmberg LP, Rytilä P, Nikander K, et al. Thirteen-year follow-up of early intervention with an inhaled corticosteroid in patients with asthma. J Allergy Clin Immunol. 2009;124(6):1180–5
8. Van Wonderen KE, van der Mark LB, Mohrs J, Geskus RB, van der Wal WM, van Aalderen WM, et al. Prediction and treatment of asthma in preschool children at risk: study design and baseline data of a prospective cohort study in general practice (ARCADE). BMC Pulm Med. 2009;9:13
9. Li D, Abhadiomhen SE, Zhou D, Shen XJ, Shi L, Cui Y. Asthma prediction via affinity graph enhanced classifier: a machine learning approach based on routine blood biomarkers. Journal of Translational Medicine. 2024 Jan 24;22(1):100.
10. Sills MR, Ozkaynak M, Jang H. Predicting hospitalization of pediatric asthma patients in emergency departments using machine learning. Int J Med Inform. 2021; 151: 104468.
11. Deliu M, Yavuz TS, Sperrin M, et al. Features of asthma which provide meaningful insights for understanding the disease heterogeneity. Clin Exp Allergy. 2018;48(1):39-47
12. Su MW, Lin WC, Tsai CH, et al. Childhood asthma clusters reveal neutrophil-predominant phenotype with distinct gene expression. Allergy. 2018;73(10):2024-2032.
13. Fitzpatrick AM, Bacharier LB, Jackson DJ, et al. Heterogeneity of mild to moderate persistent asthma in children: confirmation by latent class analysis and association with 1-year outcomes. J Allergy Clin Immunol Pract. 2020;8(8):2617-2627.e4
14. Motarjem, K., Mohammadzadeh, M., and Abyar, A. Bayesian Analysis of Spatial Survival Model with Non-Gaussian Random Effect. 2018.Journal of Mathematical Sciences, 237(5), 692–701.
15. Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119–139, August 1997
16. Friedman, J. Greedy boosting approximation: agradient boosting machine.Ann.Stat. 2021; 29, 1189–1232.