Social Science and Everant Journal Humanities Journal

ISSN: 2456-2653
Volume 09 Issue 01 January 2025

https://sshjournal.com/
Impact Factor: 2024: 6.576
2023: 5.731

DOI: https://doi.org/10.18535/sshj.v9i01.1443

Asthma Disease Prediction Using Regression Tree Method

Yustisia Lisa Christi[1, a *] **| Genrawan Hoendarto**[2,b] **| Jimmy Tjen**[3,c]

[1,2,3]Widya Dharma Pontianak University, Pontianak
[a]21421486_yustisia_l_c@widyadharma.ac.id,[b]genrawan@widyadharma.ac.id,[c]jimmytjen@widyadharma.ac.id

Received ***11-12-2024***
Revised 12-12-2024
Accepted ***02-01-2025***
Published 04-01-2025

Abstract:

Asthma is a common, chronic inflammatory disorder of the airways that affects an estimated 339 million people worldwide. Diagnostic approaches for asthma usually fail in clinical practice, partly owing to the multifactorial spectrum of the disease. This study reveals a new diagnostic algorithm where regression trees along with entropy-based subset selection(E-SS) are combined for more reliable and accurate asthma diagnosis. E-SS helps to filter out the most important features from high-dimensional datasets and avoids possible overfitting of the rate up to 91.304%, which is higher than other algorithms like Bayesian Network of 83.3%. The strength of this model is that it can capture complex (non-linear) interactions efficiently between the variables and therefore would be efficient, in particular for asthma prediction. Moreover, it is a more patient-centered methodology where risk factors of each individual are targeted. The model could aid in the diagnosis and treatment of other chronic diseases outside asthma, further alleviating global health care systems.

Keywords: Asthma, Regression Tree, Entropy-based Subset Selection, Machine Learning

Introduction:

Asthma is a chronic inflammatory disease of the airways that causes recurrent and variable symptoms, breathlessness, reversible obstruction, and bronchospasm. It impacts more than 339 million people around the world and is one of the most common chronic disorders[1]. The incidence of asthma has increased, and the rise is more noteable in developed countries, which can be attributed to urbanization, environmental changes, and lifestyle factors[2]. This problem, one of the major public health issues is associated with a reduced quality of life, a significant economic burden and high expenditure on the healthcare system.

The asthma physiology is multifactorial with genetic backgrounds predestineted and environmental exposures like allergens, pollutants, and infections playing roles. They both act in concert to elicit an overzealous immune response leading to the chronic airway inflammation, hyperresponsiveness, and structural remodeling that characterize asthma[3]. Despite the progress in knowledge and management of asthma it remains a principal cause for morbidity and mortality especially in low- and middle-income countries with limited healthcare resources[4]. Globally, asthma imposes a significant economic burden in terms of both direct and indirect costs (in medical care expenditures, productivity loss, and reduced quality of life)[5].

Given the complex and heterogeneous nature of asthma, increasing attention is being directed towards leveraging advanced statistical methods and machine learning algorithms to enhance disease management. Of these, the regression tree analysis is recognized as a handy tool in asthma research. Decision trees are a type of non-linear regression modeling technique called regression trees, that can better capture intricate relationships between multiple variables[6]. In a previous study on asthma, regression trees were used and features that best predicted disease progression and optimized treatment were identified[7].

Regression tree, implementation summary, using regression tree, researchers are able to analyze big data and detect non-linear relationships among genetic factors, environmental elements, and potential clinical phenotypes. This has resulted in the identification of different asthma phenotypes[8]. Understanding these phenotypes enables clinicians to design individualized plans for treatment that address the specific requirements of each patient and help reduce both the level of asthma-based outcomes and its overall economic cost[9]. In addition, regression tree analysis has identified biomarkers, and environmental triggers associated with severe asthma facilitating earlier intervention and targeted therapies[10].

The capacity to handle high-dimensional data and complex variable interactions is among the major merits of regression tree analysis in asthma research. Therefore, it is especially suited for implementation in the age of big data, when large-scale studies and electronic health records are increasingly being used to inform clinical practice[11]. A regression tree can enlighten the polygenic data for application in defining multiple pathways and for establishing novel therapeutic approaches[12].

Among asthma research, the usage of regression tree methods stands as an important shift in understanding and treatment for this intricate disease[13]. Over recent methods of statistical and computational analysis, the use of machine learning (ML) to healthcare data is being increasingly appreciated as a tool that would enable better understanding of this heterogeneity of asthma and predicting its progression[14]. Global Chronic Respiratory Initiative (GCRI) has the potential to revolutionize how we can manage asthma, changing the course for +300 million patients worldwide[15].

Although regression tree based models are powerful end tools and able to deal with complex and non-linear relationships between the variables, proper selection of your challenge may help in improved predictive accuracy. E-SS has been employed in a number of previous studies for feature selection in a variety of domains[16]. The standard found in this research paper has similarly been shown to outperform other popular feature selection methods, such as Principal Component Analysis(PCA). PCA consists of an orthogonal projection of a dataset to a lower dimensional space such that the variance is maximized. PCA is also used for sensor fault identification by reconstructing each variable using iterative substitution and optimization[17]. E-SS, which used feature value balance and informational value between variables to select gens by means of entropy. However, PCA cannot model non-linear dependences, and E-SS improves prediction power of the models by capturing more complex patterns hidden in data. And by removing redundant features, it cuts down on computational resources usage, a much more familiarity for high-dimensional data.

Methodology:

This section will discuss the algorithm and research flow used to develop a prediction model for asthma based on regression trees and entropy-based subset selection (E-SS). The regression tree technique focuses primarily on the study of non-linear relationships in many variables within large datasets. This would be more adequate for use when wanting to find subgroups that would show intricate patterns of variables, like in the assessment of asthma phenotypes. E-SS gives more emphasis on the feature selection part and tries to bring in the simplicity of the model by the selection of informative parameters based on entropy. For a deeper understanding of regression trees and E-SS, please refer to[18]. Based on the understanding of asthma, several parameters in the dataset can be indicators of risk or factors influencing the likehood of a person having asthma is family history of asthma, history of allergies and eczema, pollution and pollen exposure, education level, and smoking history.

Let $H = [y\ X\ D]; H \in \mathbb{R}^{m\times(n_1+n_2+1)}$ be the dataset related to asthma diagnosis, where $y \in \{0,1\}^m$ indicates whether someone has asthma (e.g., $y = 0$ means the patient does not have asthma, and vice versa), $X = [x_1\ x_2 \dots . x_{n+1}]; X \in \mathbb{R}^{m\times n_1}$ is a collection of numeric parameters (e.g., FEV1, PEF, etc), and $D \in \mathbb{Z}^{m\times n_2}$ is a collection of discrete parameters (e.g., gender, smoking history). m is the number of samples, n_1 is the number of numeric parameters, and n_2 is the number of discrete parameters.

First step: $\boldsymbol{H}$ is used to create a prediction model using only the Regression Tree. The result of this model is then compared to proposed method, which is Regression Tree with Entropy-based Subset Selection, to determine if the proposed method has higher accuracy. Let $\boldsymbol{H_v} = [y_v X_v D_v]$; $H_v \in \mathbb{R}^{m\times(n_1+n_2+1)}$ be a set of validation data similar to $\boldsymbol{H}$, but not used to train the model. Let $\%\mathrm{A}_i$ be the model prediction accuracy precentage Eq. (1):

$$\%A_i = \frac{b}{m_2} \times 100\% \qquad (1)$$

Where b_i is the number of true diagnoses based on Regression Tree-only prediction model, and m_2 is the number of samples in the validation dataset $\boldsymbol{H}_v$.

Second step: In this step, the E-SS algorithm is used to determine parameters that correlate based on entropy. Let $D_i = [x_i x_a x_{a+1} \dots . x_{a^*}]; D_i \in \mathbb{R}^{m\times a^*}; i = 1,2,3,\dots,n_1$ be the set of X is the a-th parameter from X that contains numeric parameters which have correlations based on entropy. Here, $x_a = [x_a(1)x_a(2) \dots . x_a(m)], x_a \in X$ is the a-th parameter from X. In this step, there will be $n_1 - 1$ subsets with correlations based on entropy. Specifically, each subset $\boldsymbol{H_i} = [y\ D\ D_i]$; $\boldsymbol{H_i} \in \mathbb{R}^{m\times(a^*+n_2+1)}$, is an augmented version about the patient.

Third step: A model will be built using the Regression Tree for each data subset in Eq. (2). For each, $\boldsymbol{H}_i$, suppose p_i is a Regression tree model built based on the parameters in $\boldsymbol{H}_i$. Specifically:

$$p_1 : y = F_{RT}(D, D_1), p_2 : y = F_{RT}(D, D_2), \dots . p_i : y = F_{RT}(D, D_i). \qquad (2)$$

Where F_{RT} represents the Regression Tree function[14].

Fourth step: The accuracy of each model is calculated, and the best prediction model is selected based on the highest accuracy Eq. (3):

$$\%A_i = \frac{b_i}{m_2} \times 100\% \qquad (3)$$

Where b_i is the number of correct diagnoses based on the i-th prediction model, and m_2 is the number of samples in the test dataset. The best prediction model p^* is determined using Eq. (4):

$$p^* \coloneqq \{p_i : \underset{i}{\operatorname{argmax}} \%A_i\}$$

Where p_i is the i-th model among all the prediction models. In other words $\underset{i}{\operatorname{argmax}}$ is an operator that tries to find the value of i which maximizes the function or value within argument. $\%A_i$ represents the accuracy percentage of model or algorithm i, so we want to find the algorithm i that has highest accuracy.

The parameters in D^* (the numeric parameters from p^*) will be chosen as the most significant factors in determining whether someone has asthma. The algorithm for identifying the numeric parameters used with entropy-based Subset Selection is detailed in Algorithm 1.

```
Algorithm 1: Identifying Parameters Based on Entropy

Input: Dataset described by matrix $X \in \mathbb{R}^{m \times n}$, variable to predict $j \in n$, cardinality from the subset $n^* < n$.
Output: Set of index $S \subset n, |S| = n^*$.
Initialization: $S \coloneqq \{j\} n_s = \{i \subset n | r_{ij}^2 \geq r_{min}^2\}$
Process:
    1. For $k = 1 : \min(n^* - 1, n_s)$ do
        • $j^* = argmin_{j \in n_s \setminus s} H(Z_{ij}|\beta_{ij})$
        • $S = S \cup \{j^*\}$
    2. End for
```

Where argmin refers to selecting the value of j that minimizes the function or expression $H(Z_{ij}|\beta_{ij})$. $H(Z_{ij}|\beta_{ij})$ is some function (likely conditional entropy or some other metric) that depends on j. The terms argmin actually refers to the argument minimum explain where the minimum is located inside the dataset (i.e., which sample contains the least value from the dataset).

Experimental setup:

Dataset: The data used is a dataset provided by the website kaggle. Please check this link to view it https://www.kaggle.com/datasets/rabieelkharoua/asthma-disease-dataset.

The asthma disease dataset on Kaggle, provided by Rabie Elkharoua, contains various features related to asthma patients who can further analyze the prediction and classification tasks. It is in this dataset that the focus lies because it is based on a chronic respiratory condition. Of the features for diagnosing and managing asthma, 8 features in total are considered. The following are the features in this relatively small dataset that may help in determining the lifestyle risk of developing asthma using a combination of environmental and genetic factors like age, sex, air pollution, alcohol use, and dust allergy. Occupational hazards and genetic risk further along with asthma. What makes it even smaller is that it can possibly compare the majority of its samples with so little information while still being detailed enough for initial attempts at classification.

Parameters: The number of parameters in the dataset is 29, which are: Patient ID, Age, Gender, Ethnicity, Education Level, BMI, Smoking, Physical Activity, Diet Quality, Sleep Quality, Pollution Exposure, Pollen Exposure, Dust Exposure, Pet Allergy, Family History Asthma, History Of Allergies, Eczema, Hay Fever, Gastroesophageal Reflux, Lung Function FEV1, Lung Function FVC, Wheezing, Shortness Of Breath, Chest Tightness, Coughing, Nigthtime Symptoms, Exercise Induces, Diagnosis, and Doctor In Charge.

Pre-Processing: In this section, Patient ID and Doctor In Charge were removed because they don't contribute to the determination of asthma.

Model Predictive Accuracy:

In an effort to enhance the predictive accuracy of asthma diagnosis, this study utilizes various parameters associated with with asthma risk factors and environmental variables. The table presented in the image is titled "Parameters Representation as a Variable" and serve as a representation of various parameters used in a predictive model. The table contains two main columns. First column is parameters name where the column used in the predictive model and second column is variable representation where this column present the representation of each parameter in the form of variables. These variables are used in the model to compute or analyze the relationship between these parameters and the desired predictive outcome.

Result and Discussion:

Table 1. Parameters Representation as a Variable

Parameters Name	Variable Representation	Parameters Name	Variable Representation
Age	x_1	Pollen Exposure	x_{11}
Gender	x_2	Dust Exposure	x_{12}
Ethnicity	x_3	Pet Allergy	x_{13}
Education Level	x_4	Family History Asthma	x_{14}
Smoking	x_6	Hay Fever	x_{17}
Physical Activity	x_7	Lung Function FEV1	x_{19}
Diet Quality	x_8	Lung Function FVC	x_{20}
Pollution Exposure	x_{10}		

In this section, we will discuss about True Positive Rate (TPR) Eq. (5), True Negative Rate (TNR) Eq. (6), False Positive Rate (FPR) Eq. (7), False Negative Rate (FNR) Eq. (8) and Accuracy(%A) Eq. (9):

$$TPR = \frac{TP}{TP + FN} \times 100\% \qquad (5)$$

$$TNR = \frac{TN}{TN + FP} \times 100\% \qquad (6)$$

$$FPR = \frac{FP}{TN + FP} \times 100\% \qquad (7)$$

$$FNR = \frac{FN}{FN + TP} \times 100\% \qquad (8)$$

$$\%A = \frac{TP + TN}{TP + TN + FP + FN} \qquad (9)$$

With TP representing True Positive, TN representing True Negative, FP representing False Positive, and FN representing False Negative.

Table 2. Performance Metrics of Various Subsets for Asthma Prediction Model

Subset	%A	TPR	TNR	FPR	FNR
$x_3, x_2, x_1, x_{19}, x_{17}, x_7, x_4$	88.796	0.028	0.940	0.059	0.971
$x_2, x_4, x_1, x_8, x_3, x_7, x_6, x_{20}, x_{14}, x_{10}, x_{11}$	90.886	0.028	0.962	0.037	0.917
$x_2, x_1, x_7, x_{14}, x_4, x_6, x_3, x_{13}, x_8, x_{17}$	91.304	0.057	0.965	0.034	0.942
$x_1, x_4, x_2, x_8, x_3, x_7, x_6, x_{20}, x_{14}, x_{19}, x_{11}, x_{10}$	90.886	0.028	0.962	0.037	0.971
$x_2, x_3, x_1, x_{19}, x_{17}, x_7, x_{14}, x_4, x_8, x_{27}, x_6, x_{12}$	90.050	0.072	0.951	0.048	0.927
$x_1, x_2, x_7, x_{14}, x_6, x_3, x_{14}, x_{13}, x_8, x_{17}$	91.137	0.057	0.963	0.036	0.942
$x_3, x_1, x_2, x_{19}, x_{17}, x_7, x_4$	88.628	0.028	0.938	0.061	0.971
$x_7, x_1, x_3, x_{19}, x_{17}, x_2, x_4, x_{19}, x_{12}$	89.882	0.057	0.950	0.049	0.942
$x_7, x_3, x_4, x_{19}, x_{17}, x_9, x_2, x_{16}, x_{11}$	89.715	0.028	0.950	0.049	0.971

Based on Table 2, shows various subsets of features (denoted as x_i) were used to evaluate the asthma prediction model. These subsets were obtained from the result of the regression tree enhanced by entropy-based subset selection, where the subsets represent the parameters used to understand asthma severity. Among these, a subset consisting of ten selected features (such as $x_2, x_1, x_7, x_{14}, x_4, x_6, x_3, x_{13}, x_8, x_{17}$) and highlighted in red has the highest accuracy Fig. (1). The way to validate the model is by examining several variables used in the regression tree method and calculating the percentages of TPR, TNR, FPR, and FNR, resulting in an accuracy of 91.304%.

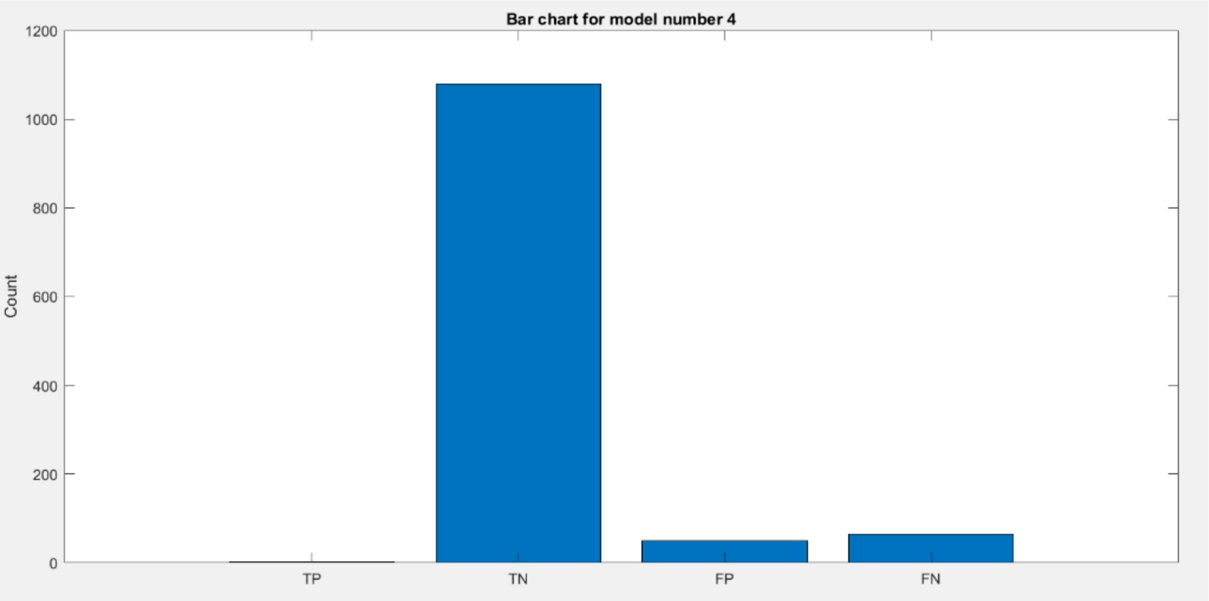

Figure 1. Bar Chart from $x_2, x_1, x_7, x_{14}, x_4, x_6, x_3, x_{13}, x_8, x_{17}$

Previous research utilizing a Bayesian network model achieved an accuracy of 83.3% in predicting asthma within an independent group of patients[19]. However, this accuracy is lower compared to the regression tree method. The regression tree method is particularly effective in handling complex and non-linear data, making it well-suited for asthma detection. It also provides string interpretability, enabling a clear understanding of how and why certain predictions are made by unveiling the underlying decision making processes. This method can effectively incorporate various factors influencing asthma, such as age, allergies, and environmental exposure.

Subset Selection

This section explains the names of indices used in the paper, as shown in the following Table 3 :

Table 3. Index Name

Index 2	2 : Age	8 : Physical Activity	5 : Education Level	4 : Ethnicity
Index 3	3 : Gender	5 : Education Level	9 : Diet Quality	8 : Physical Activity
Index 4	4 : Ethnicity	7 : Smoking	2 : Age	10 : Sleep Quality
Index 5	5 : Education Level	3 : Gender	9 : Diet Quality	8 : Physical Activity
Index 6	6 : BMI	5 : Education Level	8 : Physical Activity	10 : Sleep Quality
Index 7	7 : Smoking	4 : Ethnicity	2 : Age	10 : Sleep Quality
Index 8	8 : Physical Activity	2 : Age	5 : Education Level	4 : Ethnicity
Index 9	9 : Diet Quality	6 : BMI	8 : Physical Activity	3 : Gender
Index 10	10 : Sleep Quality	3 : Gender	9 : Diet Quality	8 : Physical Activity

Conclusion:

In conclusion, use of the regression tree analysis has been a major application in asthma research. The model has 91.304% accuracy and surpasses similar methodologies such as the Bayesian network by 83.3%. The converse implication of this work is that the model can handle non-linearity in data and complex attribute interactions, which enables a more multifaceted diagnosis of asthma based on many attributes such as age, allergies, and environmental exposure. In addition to this, this method has a powerful explanation capability, which can provide us with more clear intuitions about both how and why the predictions are made, since it is very important in some cases, such as chronic disease (asthma). This demonstrates that a regression tree can provide more accurate and relevant results in detecting asthma. This study makes a distinctive contribution to more accurate and efficient asthma detection. This research sheds light on how machine learning techniques might have a huge impact on chronic care management and healthcare in general.

References:

[1] W. H. Organization, "Chronic Respiratory," *https://www.who.int/news-room/fact-sheets/detail/asthma*, pp. 12–36, 2024.

[2] M. A, L, Hwang, "Environmental and LifeStyle Factors in Asthma," *J. Asthma Res.*, vol. 58, no. 3, pp. 231–244, 2023.

[3] S. A, M, P and R. J, Brown, "Pathogenesis of Asthma: An Overview," *Allergy Clin. Immunol. Rev.*, vol. 58, no. 3, pp. 150–165, 2022.

[4] S. Jhonson, "Asthma in Low and Middle-Income Countries: Challenge and Solutions," *Glob. Heal. J.*, vol. 10, no. 4, pp. 407–420, 2021.

[5] J. L, Ramirez and C. E, Thompson, "Economic Burden of Asthma: A Comprehensive Review," *Health Econ. Rev.*, vol. 13, no. 1, pp. 75–85, 2023.

[6] L. G, Lee, "Regression Tree Analysis: Techniques and Applications," *Stat. Methods Med. Res.*, vol. 30, no. 1, pp. 12–25, 2022.

[7] E. K, White and M. S, Zhang, "Using Regression Trees to Understand Asthma Risk Factors," *J. Clin. Epidemology*, vol. 89, pp. 45–55, 2024.

[8] T. C, Miller, "Identifying Asthma Phenotypes with Regression Trees," *Am. J. Respir. Crit. Care Med.*, vol. 207, no. 4, pp. 484–493, 2023.

[9] B. N, Kim, "Personalized Treatment Plans for Asthma: The Role of Regression Trees," *Eur. Respir. Rev.*, vol. 32, no. 167, pp. 220–230, 2023.

[10] R. J, Allen and M. P, Davis, "Biomarkers and Environmental Triggers of Severe Asthma," *Clin. Immunol.*, vol. 208, pp. 13–21, 2022.

[11] H. J, Stevens, "Handling High-Dimensional Data in Asthma Research," *J. Comput. Biol.*, vol. 30, no. 5, pp. 789–800, 2024.

[12] K. E, Robinson, "Integrating Genetic and Environmental Data in Asthma," *Nat. Rev. Genet.*, vol. 25, no. 8, pp. 532–546, 2023.

[13] P. J, Patel and R. T, Kumar, "Advancements in Asthma Research with Regression Trees," *Med. Data Anal.*, vol. 14, no. 2, pp. 102–115, 2024.

[14] L. H, Zhang, "Machine Learning in Healthcare: Applications to Asthma," *J. Biomed. Inform.*, vol. 112, pp. 103–116, 2024.

[15] N. P, Clark, "Transforming Asthma Management with Personalized Approach," *Lancet Respir. Med.*, vol. 12, no. 7, pp. 678–690, 2023.

[16] Y. Zhang and X. Zhao, "A comprehensive review of ensmble-based feature selection methods," *J. Mach. Learn.*, vol. 20, no. 1, pp. 1–25, 2019.

[17] F. Smarra and A. D. Innocenzo, “Learning methods for structural damage detection via entropy-based sensors selection,” no. September 2021, pp. 6035–6067, 2022, doi: 10.1002/rnc.6124.

[18] T. Jimmy, “Identifikasi Parameter Kualitas Bahan Pangan dengan Metode Entropy-Based Subset Selection (E-SS) (Studi Kasus: Minuman Anggur),” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1, pp. 47–54, 2024.

[19] P. He *et al.*, “Early prediction of pediatric asthma in the Canadian Healthy Infant Longitudinal Development (CHILD) birth cohort using machine learning,” no. May 2023, 2024, doi: 10.1038/s41390-023-02988-2.