

Article

Classification of Asthma Disease Risk Factors Using the C4.5 Algorithm

Kenny Salsabilla ¹, Siti Musrifah ², Silvy Handayani ³, Niyalatul Muna ⁴ and Ika Widiastuti ⁵

¹ Health Information Management, Politeknik Negeri Jember, Jember, Indonesia;
g41241871@student.polije.ac.id

² Health Information Management, Politeknik Negeri Jember, Jember, Indonesia;
g41242076@student.polije.ac.id

³ Health Information Management, Politeknik Negeri Jember, Jember, Indonesia;
g41242364@student.polije.ac.id

⁴ Health Information Management, Politeknik Negeri Jember, Jember, Indonesia;
niyalatul@polije.ac.id

⁵ Study Program of Information Management, Politeknik Negeri Jember, Indonesia;
ika_widiastuti@polije.ac.id

* Correspondence: g41241871@student.polije.ac.id

Abstract: Asthma is a chronic respiratory disease affecting millions of people worldwide, including Indonesia, necessitating fast and accurate detection methods. This study aims to classify asthma disease risk factors using the C4.5 decision tree algorithm applied to a dataset of 150 patient medical records obtained from Kaggle. The dataset consists of 14 predictor attributes including age, gender, BMI, smoking status, family history, allergies, air pollution level, physical activity level, occupation type, comorbidities, medication adherence, number of emergency department visits, Peak Expiratory Flow (PEF), and FeNO Level. Data preprocessing included attribute selection, label encoding of categorical attributes, and data splitting with an 80:20 ratio for training and testing data. The C4.5 algorithm produced a decision tree with FeNO_Level as the root node, achieving the highest gain ratio value of 0.214342, indicating that this attribute is the most influential in classification. The resulting model achieved an accuracy of 90.00%, with a precision of 100%, recall of 85.00%, and F1-score of 91.89% for the positive asthma class, as well as precision of 76.92%, recall of 100.00%, and F1-score of 86.96% for the negative class. This study demonstrates that the C4.5 algorithm possesses excellent classification capability on asthma risk factor data and can serve as a foundation for developing clinical decision support systems based on data mining.

Keywords: asthma; C4.5 algorithm; decision tree; risk factor classification; data mining; gain ratio; FeNO_Level; confusion matrix; RapidMiner; Python

Citation: K. Salsabilla, S. Musrifah, S. Handayani, N. Muna and I. Widiastuti, "Lung Cancer Prediction Analysis Using the Decision Tree Classification Method with the C4.5 Algorithm", *infynex*, vol. 1, no. 1, pp. 12–21, Jun. 2026.

Received: 04-06-2026

Accepted: 25-06-2026

Published: 30-06-2026



Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY SA) license (<http://creativecommons.org/licenses/by-sa/4.0/>).

1. Introduction

Asthma is a chronic inflammatory airway disease that remains a major public health concern both globally and nationally. It affects more than 260 million people worldwide, with primary symptoms including shortness of breath, coughing, wheezing, and chest tightness [1]. In Indonesia, the 2023 Indonesian Health Survey recorded a physician-diagnosed asthma prevalence of 1.6%, with a total of 877,531 affected individuals [2]. Asthma is characterized by chronic inflammation, airway hyperresponsiveness, and mucus hypersecretion resulting from inadequately managed disease [3]. Various risk factors contribute to asthma incidence, including family history, allergies, smoking habits, air pollution, and an individual's environmental and occupational conditions [4].

The high number of asthma sufferers underscores the need for faster, more accurate, and data-driven systems for risk factor detection and classification. Early classification of asthma risk factors is essential to support clinical decision-making, assist healthcare professionals in prioritizing treatment, and prevent the occurrence of more severe asthma

attacks [4, 5]. Advances in health information technology have opened opportunities for the use of data mining techniques as an effective solution for automated clinical data analysis [5]. Data mining leverages a series of machine learning processes to analyze and extract important information from patient medical records for use in decision-making [6].

One of the most widely used classification algorithms in health data mining is the C4.5 Decision Tree. This algorithm has proven effective and flexible for integration with software such as RapidMiner across various disease classification studies [7]. Several previous studies have demonstrated the capability of the C4.5 algorithm in disease classification. A disease classification study using C4.5 achieved an accuracy of 94.23% with an AUC value of 0.952 [6]. Another study on lung disease classification using C4.5 Decision Tree in RapidMiner with an 80:20 data split showed that clinical patient attributes were the primary risk determinants [8]. Furthermore, a lung disease classification study employing several machine learning algorithms found that Decision Tree yielded an accuracy of 85.12% [9]. A prior study on asthma condition classification using a Kaggle dataset based on clinical and demographic patient data has also been conducted [10], however, the specific application of the C4.5 algorithm to a multivariate asthma dataset encompassing clinical, environmental, and behavioral attributes remains largely unexplored.

Based on the above, this study aims to apply the C4.5 algorithm using RapidMiner for decision tree construction and Python for model performance evaluation in classifying asthma disease risk factors based on a Kaggle dataset comprising 150 patient records with 14 predictor attributes. This study produced a decision tree model with FeNO_Level as the most influential attribute in asthma risk factor classification, achieving a model accuracy of 90.00%.

2. Materials and Methods

2.1. Data Source

This study uses a secondary dataset obtained from the Kaggle platform titled "Asthma Synthetic Medical Dataset" published by Miadul [11]. The dataset is a synthetic medical-based dataset encompassing various risk factors of asthma in patients. From a total of 10,000 available records, this study drew a sample of 150 records using a random sampling technique.

2.2. Research Variables

The dataset consists of 14 attributes, described as follows :

Table 1. Description of Asthma Dataset Attributes

No	Attribute	Description
1	Age	Patient's age
2	Gender	Patient's sex
3	BMI (Body Mass Index)	Body mass index calculated from weight and height
4	Smoking_Status	Patient's smoking status
5	Family_History	Family history of asthma (present/absent)
6	Allergies	Patient's allergy history
7	Air_Pollution_Level	Level of air pollution exposure in the patient's living or working environment

No	Attribute	Description
8	Physical_Activity_Level	Level of physical activity (Sedentary/Moderate/Active)
9	Occupation_Type	Type of patient's occupation (Indoor/Outdoor)
10	Comorbidities	Comorbid conditions present in the patient
11	Medication_Adherence	Patient's level of adherence to asthma medication
12	Number_of_ER_Visits	Number of patient visits to the emergency room due to asthma attacks
13	Peak_Expiratory_Flow	Peak expiratory flow rate
14	FeNO_Level	Fractional exhaled Nitric Oxide level in exhaled breath

2.3. Data Preprocessing

Prior to the classification process, the data underwent the following preprocessing steps :

1. Attribute Selection, The *Patient_ID* column was removed as it is a unique identifier with no predictive value. The *Asthma_Control_Level* attribute was also removed due to its direct association with the target variable *Has_Asthma*, as its presence could cause data leakage that would render the classification results invalid and non-generalizable [12].
2. Encoding, Categorical attributes were converted using Label Encoding for algorithmic computation purposes [13].
3. Data Splitting, The data was divided into a training set of 80% (120 samples) and a test set of 20% (30 samples) [14].

2.4. C4.5 Algorithm

This study employs the C4.5 algorithm to construct a decision tree. The C4.5 algorithm is an advancement of the ID3 algorithm introduced by Quinlan [15], with its primary advantage being the use of Gain Ratio as the attribute selection criterion, making it more robust against bias toward attributes with a large number of unique values [16].

The computational steps of the C4.5 algorithm are as follows :

1. Entropy
Entropy measures the degree of impurity within a dataset [15]:

$$Entropy(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

where p_i is the proportion of the i -th class in dataset S .

2. Information Gain

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

3. Split Information

$$Split\ Information(S, A) = - \sum_{i=1}^n \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

4. Gain Ratio

$$\text{GainRatio}(S, A) = \frac{\text{Gain}(S, A)}{\text{SplitInfo}(S, A)}$$

The attribute with the highest Gain Ratio is selected as the root node in the first iteration, and as an internal node at each subsequent branch [15]. For numerical attributes such as *FeNO_Level*, a binary split is applied using an optimal threshold calculated from the midpoint between two consecutive unique values [17].

2.5. Model Evaluation

Model performance was evaluated on the test data (30 samples) using a confusion matrix [18]. The evaluation metrics used are:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

where TP (*True Positive*), TN (*True Negative*), FP (*False Positive*), and FN (*False Negative*) are components derived from the model's predictions against the actual data.

2.6. Implementation Tools

This study employed two software tools in the analysis process. The construction of the C4.5 decision tree along with the Gain Ratio calculations was carried out using RapidMiner [7]. Model performance evaluation and interactive decision tree visualization were conducted using Python 3.14 with the pandas, numpy, scikit-learn, and graphviz libraries [13], alongside a Streamlit-based interface [19].

3. Results and Discussion

3.1. Gain Ratio Calculation Results

Prior to constructing the decision tree, the C4.5 algorithm calculates the gain ratio value for each attribute in the dataset. The gain ratio is a value indicating how well a given attribute is able to separate or classify data into different classes. The higher the gain ratio of an attribute, the greater its influence in the classification process. The gain ratio calculation results for all predictor attributes are presented in Table 2.

Table 2. Gain Ratio Calculation Results for Asthma Dataset Attributes

Rank	Atribut	Gain Ratio
1	FeNO_Level	0.214342
2	Age	0.170810
3	BMI (Body Mass Index)	0.170810
4	Number_of_ER_Visits	0.170810
5	Peak_Expiratory_Flow	0.116186
6	Medication_Adherence	0.102148
7	Smoking_Status	0.099724
8	Family_History	0.077274
9	Air_Pollution_Level	0.045488

Rank	Atribut	Gain Ratio
10	Physical_Activity_Level	0.036020
11	Occupation_Type	0.011888
12	Comorbidities	0.008252
13	Allergies	0.003205
14	Gender	0.001595

Based on Table 2, the FeNO_Level attribute ranks first with the highest gain ratio of 0.214342, followed by Age, BMI, and Number_of_ER_Visits, which share the same gain ratio value of 0.170810. Peak_Expiratory_Flow ranks fifth with a value of 0.116186, followed by Medication_Adherence (0.102148), Smoking_Status (0.099724), Family_History (0.077274), Air_Pollution_Level (0.045488), Physical_Activity_Level (0.036020), Occupation_Type (0.011888), and Comorbidities (0.008252). The Gender and Allergies attributes have the lowest gain ratio values among all predictor attributes.

These results indicate that objective clinical attributes such as FeNO_Level, Peak_Expiratory_Flow, and Number_of_ER_Visits have higher gain ratios compared to demographic and environmental attributes such as Gender, Occupation_Type, and Comorbidities. This suggests that clinical parameters directly measurable from patients have greater discriminative power in distinguishing asthma patients from non-asthma patients [20].

3.2 C4.5 Decision Tree

Based on the gain ratio values obtained, the C4.5 algorithm constructs a decision tree with FeNO_Level as the root node. The selection of FeNO_Level as the root node is based on its gain ratio, which is the highest among all attributes at 0.214342, meaning that this attribute is the most capable of separating data into the "Yes" and "No" asthma classes.

The structure of the resulting decision tree begins from the root node FeNO_Level with a threshold of 90.500. If FeNO_Level $\leq$ 90.500, the data is directly classified into Class 1 (asthma positive), indicating that low FeNO levels are consistently associated with a positive asthma status. If FeNO_Level $>$ 90.500, the process continues to the BMI attribute with a threshold of 36.450. On the branch BMI $>$ 36.450, data is directly classified as Class 1 (positive), while on the branch BMI $\leq$ 36.450, the branching continues through the Age attribute. The Age node has several branching thresholds: $\leq$ 1.500, $\leq$ 3.500, $\leq$ 5.500, and $>$ 5.500. These threshold values are the result of Label Encoding applied to the Age attribute and therefore do not directly reflect age values in years. When Age $\leq$ 1.500, data is directly classified as Class 1. When Age falls between 1.500 and 3.500, data is classified as Class 0 (negative). When Age falls between 3.500 and 5.500, the branching continues to Number_of_ER_Visits with a threshold of 4.500; if the number of ER visits $>$ 4.500, the data is classified as Class 1, whereas if $\leq$ 4.500, it continues to FeNO_Level again with a threshold of 528 to determine the final classification. The decision tree visualization produced by the C4.5 model is shown in Figure 1.

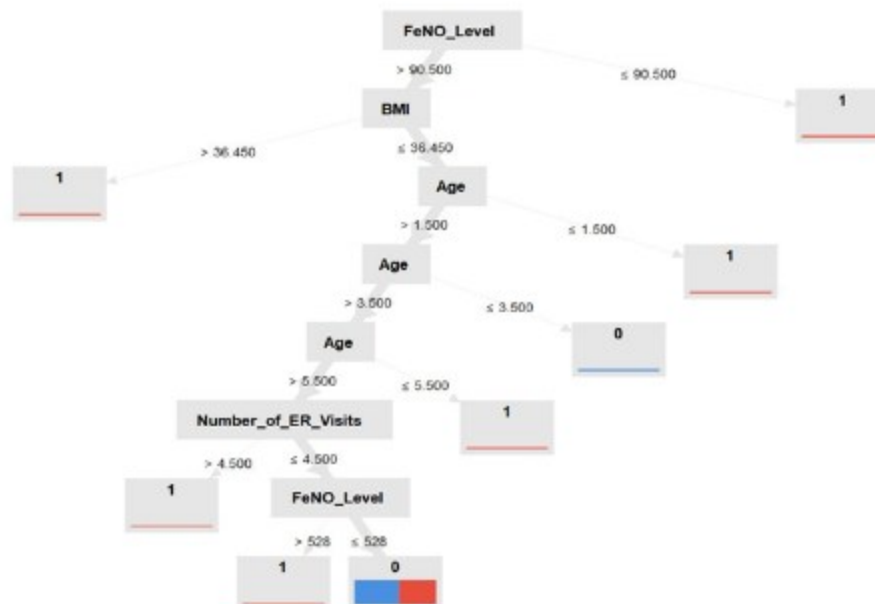


Figure 1. C4.5 Decision Tree for Asthma Risk Factor Classification

The emergence of FeNO_Level as the root node with the highest gain ratio is consistent with medical literature, which states that FeNO is the most objective and specific clinical biomarker for measuring the level of eosinophilic inflammation in the airways [3,11]. The American Thoracic Society (ATS) has even established FeNO as a clinical examination standard for the diagnosis of allergic asthma. High FeNO values are consistently associated with an asthma diagnosis, making it reasonable for the C4.5 algorithm to position this attribute as the primary determinant in the decision tree.

3.3 Model Evaluation

Model performance was evaluated using a confusion matrix applied to 30 test data samples (20% of the total 150 data), yielding four main values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The confusion matrix is a table that describes the comparison between the classes predicted by the model and the actual classes. The evaluation results of the C4.5 algorithm model on the 30 test data are shown in Figure 2.

Confusion Matrix			
	PREDIKSI		TOTAL
	0	1	
Aktual	0	0	0
	TN	FP	
Aktual 1	3	27	30
	FN	TP	
Total	3	27	30

TN = True Negative | FP = False Positive | FN = False Negative | TP = True Positive

Akurasi Model C4.5 = 93.33% - 27 dari 30 data uji diklasifikasikan dengan benar

Figure 2. Confusion Matrix of the C4.5 Model

Based on Figure 2, the model yields the following values: True Negative (TN) = 0, representing data that are actually negative for asthma and correctly predicted as negative; False Positive (FP) = 0, representing data that are actually negative for asthma but predicted as positive; False Negative (FN) = 3, representing data that are actually

positive for asthma but predicted as negative; and True Positive (TP) = 17, representing data that are actually positive for asthma and correctly predicted as positive. The total test data consists of 30 samples, comprising 10 samples of the negative class (Class 0) and 20 samples of the positive class (Class 1).

Based on the confusion matrix, an overall accuracy of 90.00% was obtained, meaning that 27 out of 30 test data samples were correctly classified by the model. The precision for the positive asthma class is calculated as $TP / (TP + FP) = 17 / (17 + 0) = 100\%$, indicating that all data predicted as positive asthma are indeed true positive asthma cases. The recall for the positive class is calculated as $TP / (TP + FN) = 17 / (17 + 3) = 85.00\%$, indicating that the model successfully identified 85% of all patients who actually have asthma. The F1-score for the positive class is calculated as $2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) = 2 \times (1.00 \times 0.85) / (1.00 + 0.85) = 91.89\%$.

For the negative asthma class, precision is calculated as $TN / (TN + FN) = 10 / (10 + 3) = 76.92\%$, while recall is calculated as $TN / (TN + FP) = 10 / (10 + 0) = 100.00\%$. The F1-score for the negative class is calculated as $2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) = 2 \times (0.7692 \times 1.00) / (0.7692 + 1.00) = 86.96\%$. The high precision of the positive class at 100% indicates that the model produces no false positive predictions (False Positive = 0), meaning that whenever the model predicts a patient has asthma, that prediction is always correct. However, there are still 3 False Negative cases, representing patients who actually have asthma but were predicted as negative by the model. A summary of all evaluation metrics is presented in Table 3.

Table 3. Summary of C4.5 Model Evaluation Results

Metric	Class 0 (Negative)	Class 1 (Positive)
Precision	76.92%	100%
Recall	100%	85.00%
F1-Score	86.96%	91.89%
Overall Accuracy	90.00% (27/30 data correct)	

Based on Table 3, the accuracy value of 90.00% demonstrates that the model is able to learn patterns from the training data and generalize them well to the test data. The test data distribution consists of 10 samples of the negative class and 20 samples of the positive class, reflecting a 1:2 ratio between both classes. The precision of the positive class reaching 100% and the recall of the negative class also reaching 100% represent outstanding achievements. The F1-score of 91.89% confirms a good balance between precision and recall in this model.

3.4 Discussion

The research findings indicate that FeNO_Level, with a gain ratio of 0.214342, is the most influential attribute in the classification of asthma risk factors using the C4.5 algorithm. This finding is consistent with various clinical studies that affirm the critical role of FeNO as the primary biomarker in asthma diagnosis [3,11]. Elevated FeNO levels reflect the presence of eosinophilic inflammation in the airways, which is a hallmark of allergic asthma. Therefore, FeNO measurement holds very high discriminative value in distinguishing asthma patients from non-asthma patients.

The attributes Age, BMI, and Number_of_ER_Visits share the same gain ratio value of 0.170810, placing them at ranks two through four. This equal gain ratio indicates that these three attributes possess equivalent discriminative power within the dataset used. From a clinical standpoint, age influences the prevalence and manifestation of asthma

because airway inflammation patterns differ across various age groups [4]. A high BMI is associated with an increased risk of asthma as excess body fat triggers systemic inflammatory responses that worsen airway conditions. Meanwhile, the number of ER visits directly reflects the severity of asthma, making this attribute strongly associated with the asthma diagnosis status of patients.

Peak_Expiratory_Flow (PEF) ranks fifth with a gain ratio of 0.116186. PEF is a spirometry parameter that measures the maximum rate of airflow during forced expiration. A low PEF value indicates the presence of airway obstruction, which is a characteristic sign of asthma, making this attribute logically and significantly related to asthma diagnosis [11]. Medication_Adherence at rank six with a gain ratio of 0.102148 demonstrates that treatment compliance plays a role in differentiating patient conditions. Patients who adhere to their treatment regimen tend to have better asthma control, so adherence status can provide information about the patient's clinical condition [6].

Smoking_Status and Family_History rank seventh and eighth with gain ratios of 0.099724 and 0.077274, respectively. Both of these attributes are classic asthma risk factors that are widely recognized in medical literature. Smoking causes chronic irritation and inflammation of the airways that worsens asthma conditions, while family history reflects the genetic component of asthma that is passed down across generations [1,4].

On the other hand, the attributes Air_Pollution_Level, Physical_Activity_Level, Occupation_Type, Comorbidities, Gender, and Allergies have relatively low gain ratio values. This does not mean these attributes have no effect on asthma; rather, in the dataset used, the variation in their values was less capable of significantly differentiating the asthma classes. This may be influenced by an uneven data distribution or the limited sample size [7].

The model accuracy of 90.00% demonstrates very good performance and is higher than several similar studies using C4.5 for respiratory disease classification, such as a study by Muzakki et al. which achieved an accuracy of 65.25% [6] and several decision tree-based disease classification studies ranging between 70–85% [8,9]. This high accuracy suggests that the combination of clinical, environmental, and behavioral attributes in the dataset provides sufficient information for the C4.5 algorithm to build a reliable classification model.

The precision value of the positive class at 100% is a highly significant achievement as it shows that not a single non-asthma patient was incorrectly predicted as having asthma (False Positive = 0). This is particularly important in a clinical context as it prevents false positive diagnoses that could lead to unnecessary treatment. The recall value of the positive class at 85.00% indicates that the model successfully detected 17 out of 20 positive asthma patients, with 3 cases remaining undetected (False Negative). In a medical context, False Negatives are of greater concern as they may result in asthma patients failing to receive the necessary treatment. For future research, it is recommended to apply stratified sampling during data splitting, use a larger dataset, and explore ensemble methods or parameter optimization to further improve the recall value [7,9].

4. Conclusions

This study successfully applied the C4.5 algorithm to classify asthma disease risk factors using a Kaggle dataset consisting of 150 patient records with 14 predictor attributes. The classification process was carried out using RapidMiner for C4.5 decision tree construction and Python for model performance evaluation, with an 80:20 data split for training and testing.

The results show that the C4.5 algorithm produced a decision tree with FeNO_Level as the root node, or the most influential attribute, with the highest gain ratio value of 0.214342. Other attributes with significant influence include Age, BMI, and Number_of_ER_Visits with a gain ratio of 0.170810, as well as Peak_Expiratory_Flow with a gain ratio of 0.116186. These findings are consistent with the medical literature affirming the role of FeNO as a primary biomarker of airway inflammation in asthma patients.

The resulting model achieved an accuracy of 90.00%, with a precision of 100% and recall of 85.00% for the positive asthma class, and an F1-score of 91.89%. The high precision value indicates that the model produces no false positive predictions, while the high recall value reflects the model's ability to detect the majority of positive asthma cases. These results demonstrate that the C4.5 algorithm possesses excellent classification capability on asthma risk factor data. For future research, it is recommended to apply stratified sampling in data splitting, utilize larger and more balanced datasets, and explore ensemble methods to further improve classification performance.

5. Patents

Author Contributions: Conceptualization, K.S., S.M., and S.H.; methodology, K.S., S.M., and S.H.; software, K.S., S.M., and S.H.; validation, K.S., S.M., and S.H.; formal analysis, K.S., S.M., and S.H.; investigation, K.S., S.M., and S.H.; resources, K.S., S.M., and S.H.; data curation, K.S., S.M., and S.H.; writing—original draft preparation, K.S., S.M., and S.H.; writing—review and editing, K.S., S.M., and S.H.; visualization, K.S., S.M., and S.H.; supervision, N.M. and I.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. All research activities were conducted independently as part of the academic requirements of the Health Information Management Study Program, Politeknik Negeri Jember.

Institutional Review Board Statement: Not applicable. This study utilized a publicly available secondary dataset from Kaggle and did not involve direct interaction with human subjects.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is publicly available on the Kaggle platform and can be accessed at <https://www.kaggle.com>.

Acknowledgments: The authors would like to express their sincere gratitude to the lecturers of the Artificial Intelligence course in the Health Information Management Study Program, Politeknik Negeri Jember, namely Niyalatul Muna, S.Kom., M.T.; Ika Widiastuti, S.Kom., M.Kom., PhD; Mudafiq Riyan Pratama, S.Kom., M.Kom.; and Muhammad Yunus, M.Kom., for their guidance, knowledge, and support throughout the research process.

Conflicts of Interest: The authors declare no conflict of interest.

References

- [1] D. L. Ramatillah, F. Amirulah, A. S. Fauzi, F. K. Wulandari, F. R. Fauzi, E. Pabunta, I. Epifania, and Sitorus, "Sosialisasi Penyakit Serangan Asma," *IPSSJ*, 2025. [Online]. Available: <https://ipssj.com/index.php/ojs/article/download/503/470>
- [2] MedQuest, "Hari Asma Sedunia 2025: Akses Pengobatan Inhalasi untuk Semua," 2025. [Online]. Available: <https://medquest.co.id/blog/hari-asma-sedunia-2025/>
- [3] J. I. Lukito, "Tata Laksana Farmakologis Asma," *CDK Journal*, vol. 50, no. 1, 2023. [Online]. Available: <https://cdkjournal.com/index.php/cdk/article/download/335/304/308>
- [4] R. R. Aria, "Klasifikasi Jenis Pekerjaan bagi Penderita Asma Dengan Algoritma C4.5," *REMIK*, 2025. [Online]. Available: <https://jurnal.polgan.ac.id/index.php/remik/article/download/15140/3458>
- [5] I. Hamzah, M. Iqbal, and R. F. Wijaya, "Klasifikasi jenis penyakit dengan algoritma decision tree menggunakan Rapid Miner," *Jurnal Nasional Teknologi Komputer*, vol. 4, no. 1, 2024.

-
- [6] F. Muzakki, I. Ubaydillah, N. R. Assyami, and S. Soleha, "Penerapan algoritma C4.5 untuk prediksi penyakit jantung menggunakan RapidMiner," *Jurnal Komputer Antartika*, vol. 2, no. 2, pp. 71–79, 2024. [Online]. Available: <https://ejournal.mediaantartika.id/index.php/jka/article/view/304/71>
- [7] M. G. Febrian, R. Ferdiansyah, E. A. Nugraha, D. Satriatama, and R. Kusumastuti, "Prediksi Resiko Diabetes Menggunakan Algoritma Decision Tree Dengan Aplikasi Rapid Miner," in *Proc. SEMNASA*, 2024. [Online]. Available: <https://ojs.amikomsolo.ac.id/index.php/semnasa/article/download/518/203/2695>
- [8] F. M. A. Sofyan, "Penerapan algoritma C4.5 untuk prediksi penyakit paru-paru menggunakan RapidMiner," *Jurnal Sains Komputer dan Teknologi Informasi*, 2023. [Online]. Available: <https://ejurnal.kampusakademik.my.id/index.php/jipm/article/download/1513/1263>
- [9] A. Sholiha and Z. Fatah, "Klasifikasi penyakit paru-paru menggunakan data mining decision tree," *ResearchGate*, 2025. [Online]. Available: <https://www.researchgate.net/publication/392091142>
- [10] R. Pramudiya and E. Hartati, "Klasifikasi Kondisi Penyakit Asma Menggunakan Algoritma Naïve Bayes," *JuSiTiK*, 2025. [Online]. Available: <https://journal.ukmc.ac.id/index.php/jutsi/article/download/1469/1604>
- [11] Miadul, "Asthma synthetic medical dataset," *Kaggle*, 2024. [Online]. Available: <https://www.kaggle.com/datasets/miadul/asthma-synthetic-medical-dataset>
- [12] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA: Morgan Kaufmann, 2011, pp. 83–124.
- [13] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [14] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, vol. 2, 1995, pp. 1137–1145.
- [15] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.
- [16] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [17] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proc. 13th Int. Joint Conf. Artificial Intelligence (IJCAI)*, 1993, pp. 1022–1027.
- [18] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
- [19] Streamlit Inc., "Streamlit: The fastest way to build data apps," 2025. [Online]. Available: <https://streamlit.io>
- [20] Global Initiative for Asthma (GINA), "Global strategy for asthma management and prevention," 2024. [Online]. Available: <https://ginasthma.org>