Geographic Analysis of Risk Factors for Chronic Respiratory Non-Communicable Diseases using Machine Learning

Ayu Susilowati[1], Andi Prademon Yunus[1*]
[1]Telkom University, Banyumas, Indonesia
Email: ayususilowati@student.telkomuniversity.ac.id, andiay@telkomuniversity.ac.id
[*]Corresponding Author

Abstract

Chronic respiratory diseases (CRDs) are a significant category of non-communicable diseases (NCDs), affecting 235 million asthma patients and 64 million COPD patients globally. In Central Java, CRDs accounted for 6% of total deaths in 2023, with WHO projecting COPD as the third leading cause of death by 2030. This study employs a decision tree machine learning approach to analyze lifestyle behaviors and environmental factors, aiming to identify key CRD risk factors and map their geographic distribution. Results indicate public transportation contributes most significantly (0.3410), followed by smoking habits and NO_2 concentration. The model achieved an RMSE of 0.40 and R^2 of 0.83, reflecting high predictive accuracy. This approach provides insights and enhances healthcare access in high-risk areas.

Keywords: Non-communicable diseases, Chronic respiratory diseases, Decision tree, Geo-mapping

1. Introduction

Non-communicable diseases (NCDs) have emerged as a significant global health challenge, with far-reaching impacts on individuals, families, and societies. Among various NCDs, chronic respiratory diseases (CRDs), such as asthma and chronic obstructive pulmonary disease (COPD), hold a prominent position due to their high prevalence and morbidity rates. According to the World Health Organization (WHO), asthma affects approximately 235 million people worldwide, while COPD impacts over 64 million individuals [1] [2]. These diseases not only pose serious health risks but also impose substantial economic burdens [3].

In Indonesia, particularly in the province of Central Java, CRDs are a major cause of mortality, accounting for 6% of total deaths in 2023 [4]. WHO projects that by 2030, COPD will become the third leading cause of death globally, underscoring the urgency of addressing this issue through data-driven approaches [5]. Referring to global policies, controlling risk factors is a crucial aspect of preventing non-communicable diseases (NCDs) [6]. It is essential to conduct research to identify the risk factors for chronic respiratory diseases to prevent the rising prevalence of NCDs associated with respiratory conditions. One preventive approach involves the application of machine learning techniques.

Machine learning (ML) is becoming an increasingly important tool in health big data analysis due to its ability to identify complex patterns that are difficult to find through conventional statistical methods [7][8]. In the context of chronic respiratory disease risk factors, this study examines the distribution of public facilities—such as transportation, sports, and recreation centers, and access to healthy food—in Central Java. Using geographic analysis within a decision-tree framework, this study aims to elucidate the interactions between risk factors associated with chronic respiratory diseases in Central Java.

By analyzing feature importance, we can identify which predictors, such as air pollution levels, smoking prevalence, and access to health-related facilities, have the most significant impact on the occurrence of respiratory diseases. This approach not only aids in building predictive models with high accuracy but also provides actionable insights for policymakers and public health officials to target interventions effectively [9]. In this study, feature importance derived from a Decision Tree model is utilized to quantify the contribution of each risk factor, allowing for the identification of dominant factors at the district and city levels [10] [11].

Thus, geo-mapping distribution analysis with decision-tree methods offers a strategic approach to uncovering risk factor patterns for chronic respiratory diseases in Central Java, providing an evidence-based foundation for more effective prevention and control measures targeting these conditions.

2. Methodology

2.1. *Feature Importance Using Decision Tree*

In this subsection, we employ a machine learning approach, specifically the decision tree algorithm, to identify and analyze key risk factors contributing to the prevalence of chronic respiratory diseases (CRDs) in Central Java. The decision tree is a supervised learning model that is widely used in classification and regression

tasks. It works by splitting the dataset into subsets based on feature values, creating a tree-like structure that aids in understanding the relationships between variables and the target outcome.

The core strength of the decision tree lies in its ability to compute *feature importance*, which quantitatively measures the contribution of each feature in improving the predictive accuracy of the model [12]. To calculate *feature importance*, the decision tree uses the *gini impurity* metric to evaluate the quality of splits at each node. The *gini impurity* is a measure of the likelihood that a randomly chosen data point would be incorrectly classified if it were randomly assigned a label according to the distribution of labels in the dataset. It is calculated as follows in Eq. (1).

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \quad (1)$$

Where $\boldsymbol{p_i}$ is the proportion of samples belonging to class $\boldsymbol{i}$, and $\boldsymbol{n}$ is the total number of classes. At each split, the decision tree computes the *gini impurity* for the parent node ($\boldsymbol{Gini_{\text{parent}}}$) and the resulting child nodes ($\boldsymbol{Gini_{\text{left}}}$) and ($\boldsymbol{Gini_{\text{right}}}$). The reduction in *Gini impurity*, referred to as the *Gini Gain*, determines the quality of the split. The formula for *Gini Gain* is as follows in Eq. (2).

$$Gini\ Gain = Gini_{\text{parent}} - Gini_{\text{child}} \quad (2)$$

By systematically evaluating the *Gini gain* at each split, the decision tree identifies the feature that most effectively reduces impurity, thereby improving the model's accuracy. This process continues iteratively, building a tree structure that captures the relationships between features and the target variable while ranking feature importance based on their contribution to impurity reduction.

2.2. *Dataset and Variables*

The dataset used in this study consists of 35 observations representing data from various districts and cities in Central Java. It includes variables related to chronic respiratory diseases (CRD) and potential risk factors. Key variables are: CRD_case (number of CRD cases), cigarette (average cigarette consumption per capita), NO_2 and SO_2 (air pollutant concentrations in µg/m³), and various facilities such as transport (public transportation), gym, fast_food outlets, sweet_drink vendors, tourism_spot, alcohol_store, and healthy_food outlets.

The facility data were collected through web scraping of mapping platforms to capture their distribution across districts and cities, providing a comprehensive geographic overview. This dataset integrates environmental and lifestyle factors to analyze their influence on CRD prevalence, enabling the identification of localized risk factors and supporting data-driven policy recommendations for disease prevention and control.

2.3. *Model Evaluation and Geo-mapping*

To evaluate the decision tree model, *Root Mean Square Error (RMSE) and R-squared (R²)* were calculated to measure the model's accuracy. RMSE evaluates the average error between predicted and observed values, while R² explains the proportion of variance in the dependent variable that is predictable from the independent variables. The formulas used are as below in Eq. (3) and Eq. (4).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (4)$$

In the Eq. (3) for model evaluation, y_i represents the actual value or the observed data, $\hat{y}_i$ denotes the predicted value generated by the model, and nnn refers to the total number of observations used in the calculation. These components are crucial for assessing the accuracy of the model by comparing the predicted outcomes with the actual results across all data points. Meanwhile, Eq. (4) evaluation context, $\bar{y}_i$ represents the mean of the actual values. This serves as a baseline to measure the model's performance by comparing the predicted values with the average of the observed data. It is particularly useful in calculating R^2 metrics like the coefficient of determination to evaluate how well the model explains the variance in the actual data.

To visualize the spatial distribution of chronic respiratory disease (CRD) prevalence and its significant risk factors, a geo-mapping approach was employed using GeoJSON data. This process began with integrating data on CRD prevalence and the feature importance of key risk factors into GeoJSON files containing the administrative boundaries of districts and cities in Central Java. Subsequently, a choropleth map was generated to display the CRD prevalence, with varying color intensities representing the prevalence levels across regions. Overlay markers highlighted significant risk factors such as public transportation, cigarette consumption, and NO_2 concentrations. This visualization approach enables a clearer understanding of geographical patterns and the influence of key risk factors, providing valuable insights for targeted intervention strategies.

3. Results and Discussions

3.1. *Simulations settings*

In this section, we are using the decision tree approach as the selected machine learning method to predict chronic respiratory disease (CRD) cases. The model was initialized using the **DecisionTreeRegressor** from the

sklearn.tree library, as the target variable, CRD_case, represents continuous data. To ensure reproducibility, the parameter **random_state=42** was applied, fixing the random seed for consistent results. Model optimization was carried out through hyperparameter tuning using GridSearchCV, which systematically explored various parameter combinations to enhance performance. Key parameters included max_depth, which controls the maximum depth of the tree and the model's complexity; min_samples_split, determining the minimum number of samples needed to split a node; min_samples_leaf, specifying the minimum samples required at a leaf node; and max_features, which limits the maximum number of features considered during splits. The refined parameter combinations identified through this process resulted in an optimized decision tree model capable of effectively predicting and analyzing CRD cases.

3.2. *Performance and Analysis*

The parameter tuning process for the decision tree model was conducted using GridSearchCV, which systematically tested various parameter combinations to identify the optimal configuration. The parameters adjusted included max_depth, min_samples_split, min_samples_leaf, and max_features. Each combination was evaluated based on the model's predictive performance using metrics such as Root Mean Squared Error (RMSE) and R^2 score. The results of the parameter tuning experiments are summarized in Table 1, highlighting the best-performing parameter configuration that achieved the highest evaluation score.

Table 1. Parameter Tuning Results and Best Configuration

Max Depth	Min Samples Split	Min Samples Leaf	Max Features	RMSE	R^2
5	10	4	sqrt	0,75	0,39
10	5	8	None	0,82	0,27
15	5	2	sqrt	0,83	0,26
20	10	6	sqrt	0,85	0,22
10	5	2	$\log^2$	0,42	0,81
5	**5**	**1**	**sqrt**	**0,40**	**0,83**

The results in Table 1 as shown in the highlighted row, utilized a max_depth of 5, min_samples_split of 5, min_samples_leaf of 1, and max_features set to "sqrt," resulting in the lowest **RMSE (0.40)** and highest **R^2 score (0.83)**, which delivered the best evaluation results.

After obtaining the best model through GridSearchCV, the evaluation was conducted using cross-validation and Root Mean Squared Error (RMSE) as performance metrics. Cross-validation ensures the model's ability to generalize new data by splitting the training data into several folds, such as six folds. In each iteration, the model was trained in five folds and tested on one, rotating through all folds. This method provides a more reliable estimate of how the decision tree model would perform on unseen data, thereby reducing the risk of overfitting.

When tested on the full dataset, the trained model achieved an RMSE of 0.40 and an R^2 score of 0.83. The RMSE value of 0.40 reflects the model's low prediction error, while the R^2 score of 0.83 indicates that the model successfully explains 83% of the variance in the target variable (CRD_case). These evaluation results highlight the model's strong predictive performance, with high accuracy and a low error rate, making it a robust approach for understanding chronic respiratory disease prevalence in the given dataset.

Next, feature importance is determined to assess the contribution of each feature in predicting the target variable. Feature importance measures the impact of each feature on the data splits throughout the decision tree. These values are derived from the model selection process and indicate the relative contribution of each feature in segmenting the data. Higher importance values suggest a greater influence of the feature on the model's decision-making process. In this context, the feature importance reflects how significantly each independent variable impacts the prediction of chronic respiratory disease (CRD) cases. The greater the feature's importance, the more substantial its role in shaping the model's decisions. Visualization of feature importance is then performed to facilitate the identification of key features with the most significant contributions to the model. The following Table 2. presents the feature importance values and their corresponding visual representation.

Table 2. Feature Importance Value

Feature	Importance Value
transport	0.3410
cigarette	0.2537
NO_2	0.1310
tourism spot	0.0833
alcohol store	0.0656
SO_2	0.0403
healthy food	0.0352
fast food	0.0269
gym	0.0229
sweet drink	0.0000

From the feature importance Table 2, the visualization is presented on a map of districts and cities in Central Java using GeoJSON. This map illustrates the prevalence of chronic respiratory diseases (CRD) across regions, with color intensity representing the CRD case distribution of significant risk factors, such as public transportation facilities, cigarette consumption rates, and NO_2 concentrations in each area shown.

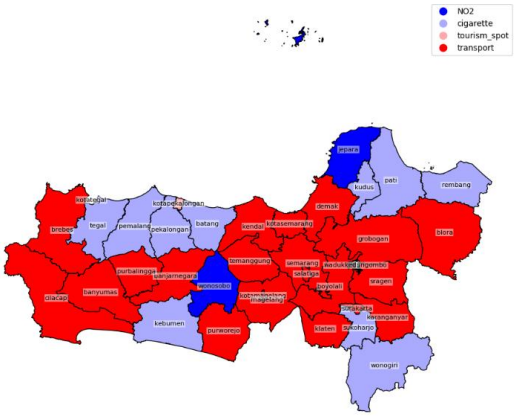

Fig.1 Feature importance of geo-mapping in Central Java

The visualization shown Fig.1 results reveal that the number of public transportation facilities (transport) has the highest feature importance value of **0.3410**, indicating that public transportation is a major risk factor for chronic respiratory diseases (CRD). The map visualization confirms that most regions in Central Java are dominated by this risk factor. This can be interpreted as the contribution of air pollution generated by public transportation vehicles, along with other motorized vehicles such as cars and motorcycles. Areas with dense public transportation facilities are often highly exposed to vehicle emissions, including nitrogen dioxide (NO_2) and fine particulate matter (PM2.5), which are concentrated in high-traffic zones, increasing the risk of respiratory diseases. Long-term exposure to these pollutants is scientifically linked to airway inflammation, decreased lung function, and a heightened risk for individuals with pre-existing chronic respiratory conditions. Furthermore, public transportation systems, particularly in urban areas, tend to have high population densities, which exacerbate exposure to secondhand smoke and other environmental allergens such as dust or mold. These factors may act as triggers that worsen the condition of individuals with chronic respiratory diseases while simultaneously elevating the risk of new cases in vulnerable populations.

Cigarette consumption (cigarette) has a significant importance value of **0.2537**. Smoking is a leading cause of respiratory disorders, including chronic obstructive pulmonary disease (COPD) and asthma. Regions such as Kudus, Pemalang, and Rembang show a dominance of cigarette smoke as a risk factor, indicating a high rate of cigarette consumption in these areas. This not only affects active smokers but also the surrounding community due to passive smoke exposure. The significance of this variable is supported by literature that highlights how the toxic chemicals in cigarette smoke directly damage lung tissue and impair respiratory capacity. This effect is not limited to active smokers but extends to individuals exposed to passive smoke, which contains carcinogenic compounds. Furthermore, children exposed to secondhand smoke are at higher risk of lung growth issues, asthma, and respiratory infections. Additionally, individuals with pre-existing conditions like COPD or asthma may experience exacerbations or worsened symptoms due to smoke exposure.

The NO_2 variable **(0.1310)** also plays a significant role in predicting chronic respiratory diseases. Regions such as Jepara and Wonosobo show NO_2 pollution as a dominant risk factor. Long-term exposure to NO_2 has been scientifically proven to impair lung function and increase the risk of asthma. NO_2 is categorized as a carcinogen and has particularly harmful effects on infants and the elderly. It originates from transportation emissions fueled by fossil fuels, industrial smoke, and environments already contaminated by NO_2. Industrial activities, especially those involving the combustion of fossil fuels such as coal, natural gas, or oil, contribute significantly to NO_2 emissions. Power plants, refineries, and manufacturing factories are examples of industries that generate NO_2.

Other variables, such as tourism spots (tourism_spot) with a value of **0.0833**, found in cities like Pekalongan, and alcohol stores (alcohol_store) with a value of 0.0656, have lower but still relevant contributions. The presence of tourist spots can influence local pollution levels due to increased mobility, while alcohol stores may serve as indicators of urbanization or risky lifestyle patterns. Variables like healthy food (healthy_food), fast food, and gyms show lower importance values (<0.04), indicating minimal impact on the risk of chronic respiratory diseases in Central Java.

4. Conclusion

In this paper, we employed a decision tree model to identify the key risk factors contributing to the prevalence of chronic non-communicable respiratory diseases in Central Java. The model calculates feature importance using Gini impurity, revealing that public transportation, cigarette consumption, and NO_2 concentration are the most significant risk factors, with public transportation having the greatest influence (0.3410). Model evaluation, with an RMSE of 0.40 and R^2 of 0.83, demonstrates strong predictive accuracy.

The visualization of risk factors across districts and cities in Central Java highlights regional variations in risk influences. The map predominantly features red areas, indicating public transportation as the leading risk factor due to air pollution from fossil fuel emissions, notably nitrogen dioxide (NO_2). Regions such as Jepara and Wonosobo show a strong correlation with NO_2 pollution (depicted in blue), reinforcing the link between air pollution and disease prevalence. Additionally, areas marked in purple, such as Kudus, Pemalang, and Rembang, suggest that smoking is the primary risk factor, both through active smoking and passive exposure to secondhand smoke. These findings emphasize the complex interplay of environmental and behavioral factors in shaping respiratory health risks across different regions.

References

1. S. C. Dharmage, J. L. Perret, and A. Custovic, "Epidemiology of asthma in children and adults," Frontiers in Pediatrics, vol. 7, p. 246, 2019, doi: 10.3389/fped.2019.00246.
2. E. Boers et al., "Global burden of chronic obstructive pulmonary disease through 2050," JAMA Network Open, vol. 6, no. 12, p. e2346598, 2023, doi: 10.1001/jamanetworkopen.2023.46598.
3. D. E. Bloom, S. Chen, M. Kuhn, M. E. McGovern, L. Oxley, and K. Prettner, "The economic burden of chronic diseases: Estimates and projections for China, Japan, and South Korea," The Journal of the Economics of Ageing, vol. 17, p. 100163, 2020.
4. H. Y. Lu, C. F. Chen, D. L. Lee, Y. J. Tsai, and P. C. Lin, "Effects of early pulmonary rehabilitation on hospitalized patients with acute exacerbation of chronic obstructive pulmonary disease: A systematic review and meta-analysis," International Journal of Chronic Obstructive Pulmonary Disease, vol. 18, pp. 881–893, 2023, doi: 10.2147/COPD.S397361.
5. C. T. Wu et al., "Acute exacerbation of chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: Development and cohort study," JMIR Mhealth and Uhealth, vol. 9, no. 5, p. e22591, 2021, doi: 10.2196/22591.
6. A. Budreviciute et al., "Management and prevention strategies for non-communicable diseases (NCDs) and their risk factors," Frontiers in Public Health, vol. 8, p. 574111, 2020, doi: 10.3389/fpubh.2020.574111.
7. A. B. Mahammad and R. Kumar, "Machine learning approach to predict asthma prevalence with decision trees," in 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022, pp. 263–267, doi: 10.1109/ICTACS56270.2022.9988210.
8. S. B. Khanagar et al., "Developments, application, and performance of artificial intelligence in dentistry: A systematic review," Journal of Dental Sciences, vol. 16, no. 1, pp. 508–522, 2021.
9. K. Poonpon et al., "Enhancing predictive accuracy in educational assessment: A comparative analysis of machine learning models for predicting student performance," Review of Contemporary Philosophy, vol. 23, pp. 142–160, 2024.
10. Y. Y. Zhou et al., "Risk factor analysis and clinical decision tree model construction for diabetic retinopathy in Western China," World Journal of Diabetes, vol. 13, no. 11, pp. 986–1000, 2022, doi: 10.4239/wjd.v13.i11.986.
11. P. Du et al., "The application of decision tree model based on clinicopathological risk factors and pre-operative MRI radiomics for predicting short-term recurrence of glioblastoma after total resection: A retrospective cohort study," American Journal of Cancer Research, vol. 13, no. 8, pp. 3449–3462, 2023.
12. J. Zhang et al., "Improving wave height prediction accuracy with deep learning," Ocean Modelling, vol. 188, p. 102312, 2024, doi: 10.1016/j.ocemod.2023.102312.

Authors Introduction

Ms. Ayu Susilowati

She is currently pursuing the Bachelor of Informatics with Honours in Faculty of Informatics, Telkom University Purwokerto, Indonesia. His research interests are Artificial Intelligence, data and bussiness analyst.

Dr. Andi Prademon Yunus

He is an Assistant Professor at Telkom University, and he received his PhD in Engineering from Mie University, Japan. His research focuses on applied and fundamental machine learning for motion and behavior computing. He also collaborates with industry partners to develop AI-based tools for language modeling and image analytics.