Exploring Non-Communicable Disease Risk Factors on Cancer Rates in Central Java Using Random Forest and SHAP

Novi Ramadani, Andi Prademon Yunus*
Telkom University, Banyumas, Indonesia
Email: novrmd@student.telkomuniversity.ac.id, andiay@telkomuniversity.ac.id
**Corresponding Author*

Abstract

Non-communicable diseases (NCDs), including cancer, account for over 75% of global deaths, with cancer being the second leading cause. In Indonesia, cancer ranks third, and Central Java has a high prevalence of 1.7 per 1,000 population in 2023. This study analyzed cancer incidence in Central Java using a Random Forest model combined with SHapley Additive exPlanations (SHAP) to assess local feature importance. Key risk factors identified include the presence of sugar-sweetened beverage outlets, which affect 28.6% of districts, air pollution from SO_2 and NO_2, which impacts 22.9% and 11.4% of districts, respectively, and the presence of slum areas, which is associated with higher cancer risks in 8.6% of districts. These findings offer insights into targeted public health strategies aimed at reducing cancer incidence and improving community health.

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

1. Introduction

Non-communicable diseases (NCDs), including heart disease, stroke, cancer, diabetes, and chronic respiratory diseases, are responsible for over 75% of global deaths. Among these, cancer stands as the second leading cause of mortality worldwide, claiming approximately 10 million lives annually [1]. In Indonesia, cancer ranks as the third leading cause of death, following cardiovascular diseases and maternal health issues [2]. Based on data from the Indonesian Ministry of Health in 2023, Central Java, in particular, is among the top five provinces with the highest cancer prevalence, recorded at 1.7 per 1,000 population.

The main risk factors for non-communicable diseases, including cancer, encompass unhealthy eating habits (such as consuming junk food), smoking, alcohol use, physical inactivity, and conditions like high blood pressure, high blood glucose, high cholesterol, and obesity. Socio-economic and environmental factors, such as poverty, limited healthcare access, low public health spending, air pollution, climate change, and sun exposure, also contribute [3], [4]. Understanding these risk factors is crucial for designing effective public health interventions. This study examines the spatial distribution and relative importance of cancer risk factors in Central Java, focusing on the presence of fast-food outlets, tourist attractions, slum areas, transportation hubs, alcohol stores, sugary drinks outlets, gyms, sports halls, smoking prevalence, and air pollution (NO_2 and SO_2).

Machine Learning (ML) has emerged as a transformative tool for analyzing complex datasets, enabling researchers to uncover non-linear relationships and patterns [5]. Among ML algorithms, Random Forest (RF) is widely regarded for its robustness and versatility. RF operates by constructing multiple decision trees and combining their outputs, improving accuracy, and mitigating overfitting. Moreover, RF provides insights into feature importance, identifying key predictors of outcomes [6][7]. In this study, RF is employed to evaluate the global importance of cancer risk factors, complemented by SHapley Additive exPlanations (SHAP) to assess local variations and spatial distributions. SHAP values explain the contribution of each feature to the model's predictions for individual data points, offering insights into the influence of features on the model's decisions for each sample [8].

This research sheds light on the spatial disparities of cancer risk factors in Central Java, offering valuable insights for tailored public health strategies. By utilizing advanced computational methods like RF and SHAP, this study demonstrates the potential of Machine Learning in addressing intricate public health challenges and supporting data-driven decision-making.

2. Material and Method

2.1. Material

2.1.1. Cancer Data

Cancer cases data were obtained from the Health Office of Central Java Province, which includes information on the incidence of cancer, covering leukemia, cervical cancer, breast cancer, retinoblastoma, and colorectal cancer in each administrative region. The administrative regions consist of 29 regencies and 6 cities

in Central Java. This data serves as the foundation for understanding the spatial distribution of cancer cases in the region.

2.1.2. Risk Factors Data

Risk factor data were collected through various methods. Information related to public facilities and community behaviors, such as sports hall, gym, tourist attractions, fast food outlets, sugary drinks outlet, alcohol store, and transportation hub, was obtained through web scraping techniques from Google Maps. Additionally, air pollution data (NO_2 and SO_2) and smoking prevalence data were collected from Central Bureau of Statistics (BPS) reports. Information regarding slum area size was obtained from Central Java Province Settlement Area Information System reports.

2.2. Method

2.2.1. Random Forest (RF)

Random Forest (RF) is a non-parametric machine learning method for classification and regression [8]. It builds multiple decision trees from randomly selected samples in the training data. The main steps of RF are:

1. Randomly select n samples from the training set with replacement (usually 2/3 of the data). The remaining third is used for out-of-bag (OOB) estimation.
2. For each sample, randomly select a subset of variables and create a decision tree.
3. Trees grow to their maximum size without pruning.
4. The prediction/classification is based on the majority vote (for classification) or average (for regression) from all trees.

The OOB method is also utilized to evaluate the significance of each independent variable. A common approach to measure importance is by calculating the increase in Mean Squared Error (%IncMSE). This technique involves randomly permuting the values of each variable in the OOB sample and then calculating the resulting OOB error. If the OOB error rises with the permuted values, it suggests that the variable plays a significant role. The greater the increase in error, the more crucial the variable is for predicting the dependent variable [9].

To evaluate the performance and goodness-of-fit of the RF model, the following common metrics are computed: Root Mean Square Error ($RMSE$) (1) and coefficient of determination (R) (2).

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

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

Where y_i is the actual value of observation for sample i, $\hat{y}_i$ is the predicted value for observation i, and $\bar{y}$ is the average value of the dependent variable, and n refers to the total sample.

2.2.2. SHAP (Shapley Additive Explanations)

SHAP (Shapley Additive Explanations) method is employed to measure feature importance at the local level within predictive models, providing clearer insights into the contribution of each feature to the model's decision-making process. In this study, SHAP can be used to evaluate how specific risk factors influence cancer risk predictions in different regions of Central Java.

SHAP generates locally additive feature attributions, meaning that the contribution of each feature to the model's prediction can be calculated and summed up to determine the difference between the actual prediction and the model's average prediction, as described in

$$\hat{y}_i = shap_0 + shap(X_{1i}) + shap(X_{2i}) + \cdots + shap(X_{pi}) \quad (3)$$

Where $shap_0 = E(\hat{y})$ is the average prediction for all observations, and $shap(X_{ji})$ represents the SHAP value of the j^{th} feature for observation i, indicating the feature's marginal contribution to the prediction. The sum of all SHAP values equals the difference between the actual prediction and the average prediction. SHAP values maintain the properties of Shapley values. Additionally, the absolute SHAP value indicates the magnitude of the feature's influence on the model's prediction, making it a useful measure of feature importance [10].

3. Results

Before fitting the RF model and to prevent overfitting, we used Grid Search to find the optimal values for the hyperparameters. After testing various combinations of hyperparameter values through K-fold cross-validation, we obtained the following settings for the RF model: The best hyperparameter configuration, determined through 3-fold cross-validation, resulted in an of $\boldsymbol{R^2}$ 0.84 and an $\boldsymbol{RMSE}$ of 0.39. The optimal configuration used the following values: n_estimators = 80, max_depth = 7, min_samples_split = 2, min_samples_leaf = 1, and max_features = sqrt.

The importance of the independent variables in the Random Forest (RF) model is illustrated in Fig. 1. Variables with a higher percentage increase in the mean squared error (%IncMSE) are considered more important. The top five variables contributing to cancer rates are SO_2, sugary drinks, smoking prevalence (%), NO_2, and slum area.

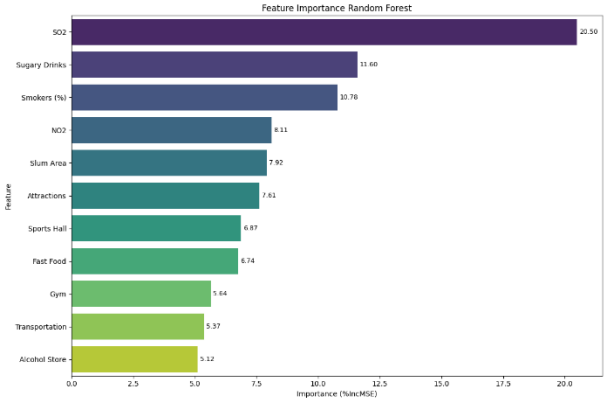

Fig. 1. RF feature importance

We also determined the proportion of counties sharing the same primary local risk factor, defined as the factor with the highest importance (Table 1). For instance, sugary drinks are the factor with the highest importance in 28.6% of counties.

Table 1. Counties sharing the same primary factor, identified as the one with the highest importance.

Local Primary Factor	Share of Counties (%)
Sugary Drinks	28.6%
SO_2	22.9%
NO_2	11.4%
Slum Area	8.6%
Attractions	8.6%
Smokers (%)	5.7%
Fast Food	5.7%
Sports Hall	5.7%
Alcohol Store	2.9%
Transportation Hub	0%
Gym	0%

Unsurprisingly, factors such as consumption habits, air pollution, and environmental conditions, including sugary drinks (28.6%), SO_2 (22.9%), NO_2 (11.4%), and slum area (8.6%), rank as the most influential factors contributing to cancer incidence in 71.5% of regions in Central Java (Table 1, Fig. 2). Additionally, attractions (8.6%), followed by smokers (5.7%), fast food outlets (5.7%), sports hall (5.7%), and alcohol stores (2.9%), are also identified as primary risk factors. The spatial distribution of these factors reveals intriguing patterns (Fig. 2).

Sugary drinks dominate as the primary factor in Banyumas, Boyolali, Brebes, Grobogan, Kebumen, Kendal, Klaten, Semarang City, Semarang, and Tegal. SO_2 emerges as the most influential risk factor in Banjarnegara, Batang, Blora, Surakarta City, Tegal City, Pemalang, Purworejo, and Rembang. NO_2 is dominant in Cilacap, Magelang, Pati, and Purbalingga, while slum areas significantly impact Karanganyar, Kota Salatiga, and Wonosobo. Meanwhile, attractions are the leading factor in Pekalongan, Sragen, and Sukoharjo. Smokers (10%) influence Demak and Wonogiri, fast food outlets dominate in Jepara and Pekalongan City, sports halls are influential in Magelang City and Kudus, and alcohol stores are significant in Temanggung.

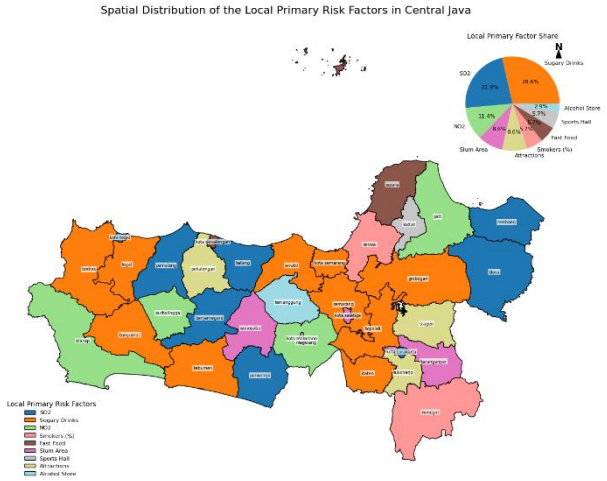

Fig. 2. Spatial distribution of key factor importance

4. Discussion

This cross-sectional ecological study analyzed cancer incidence rates at the county level in Central Java in 2023. Using a Random Forest model (R^2=0.84, RMSE=0.39), the study effectively captured complex, non-linear relationships between local risk factors and cancer incidence. It aimed to identify the key environmental, socio-economic, and lifestyle factors influencing cancer rates. Below, we explore these factors in detail and emphasize their significance.

4.1 Sugary Drinks

Sugary drinks are a primary local risk factor affecting 28.6% of Central Java's districts, including Banyumas, Boyolali, Brebes, Grobogan, Kebumen, Kendal, Klaten, Kota Semarang, Semarang, and Tegal. Long-term high sugar consumption increases the risk of non-communicable diseases like cancer, obesity, and type 2 diabetes [11]. Strategies such as reducing sugar intake, clear nutritional labeling, and public health campaigns are essential to address this issue.

4.2 SO_2

SO_2 impacts 22.9% of districts, primarily from fossil fuel combustion in power plants and vehicles. Prolonged exposure to SO_2 is linked to tissue damage and oxidative stress, increasing cancer risk [12]. Industrial areas show the highest SO_2 influence, highlighting the need for stricter emission regulations and investment in eco-friendly technologies to improve public health.

4.3 NO_2

Nitrogen dioxide (NO_2) is a dominant risk factor in Cilacap, Magelang, Pati, and Purbalingga, largely due to vehicle emissions and industrial activity. Long-term NO_2 exposure is associated with systemic inflammation and increased cancer risk [13]. Policies targeting emission reduction and better air quality monitoring systems are critical to mitigate health risks.

4.4 Slum Area

Slum areas, characterized by overcrowding, poor sanitation, and limited access to clean water [14], represent a dominant risk factor in 25.7% of districts. These areas often experience higher pollution levels and limited healthcare access, which exacerbate cancer risks. Urban development programs aimed at improving infrastructure and health services are crucial to mitigating these challenges.

5. Conclusions

This study examined cancer incidence rates in Central Java and identified key risk factors using a Random Forest model, complemented by SHapley Additive exPlanations (SHAP) for detailed local feature importance analysis. The findings revealed that local environmental, social, and lifestyle factors contribute significantly to the variation in cancer rates across regions. Sugar-sweetened beverage consumption affected 28.6% of districts, highlighting the importance of public health strategies to promote healthier dietary habits. Air pollution from SO_2 and NO_2 emissions was a major risk in 22.9% and 11.4% of districts, respectively, emphasizing the need for stricter pollution control policies and cleaner technologies. Additionally, slum areas were associated with higher cancer risks in 8.6% of districts, underscoring the importance of improving living conditions and access to healthcare.

These findings provide valuable insights for policymakers to design targeted interventions that address local risk factors, reduce cancer incidence, and promote healthier communities.

References

1. World Health Organization: WHO. (2024, December 23). *Noncommunicable diseases*. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.
2. World Health Organization: WHO. *Noncommunicable Diseases Data Portal*. https://ncdportal.org/.
3. Manderson, L., & Jewett, S. (2023). Risk, lifestyle, and non-communicable diseases of poverty. Globalization and Health, 19(1). https://doi.org/10.1186/s12992-023-00914-z.
4. Piovani, D., Nikolopoulos, G. K., & Bonovas, S. (2022). Non-Communicable Diseases: the Invisible Epidemic. *Journal of Clinical Medicine*, *11*(19), 5939. https://doi.org/10.3390/jcm11195939
5. Grekousis, G. (2018). Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis. *Computers Environment and Urban Systems*, *74*, 244–256. https://doi.org/10.1016/j.compenvurbsys.2018.10.008.
6. Breiman, L. (2001). Random forest. *Machine Learning*, *45*(1), 5-32.
7. Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, 2024, 69–79. https://doi.org/10.58496/bjml/2024/007.
8. Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, 34(10), 1013–1026. https://doi.org/10.1007/s10822-020-00314-0.
9. Georganos, S., Grippa, T., Gadiaga, A. N., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., & Kalogirou, S. (2019). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, 36(2), 121–136. https://doi.org/10.1080/10106049.2019.1595177.
10. Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers Environment and Urban Systems*, *96*, 101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845.
11. De Lorgeril, M., Salen, P., & Rabaeus, M. (2020c). Sugary drinks and cancer risk. *Translational Cancer Research*, 9(5), 3172–3176. https://doi.org/10.21037/tcr-2020-003.
12. Xu, Q., Lu, C., Murithi, R. G., & Cao, L. (2021b). Increase associated risk of gynaecological cancer due to long-term exposure to high concentration of atmospheric SO_2 industrial pollutant. *Indoor and Built Environment*, 31(8), 2183–2192. https://doi.org/10.1177/1420326x211003655.
13. Amadou, A., Praud, D., Coudon, T., Deygas, F., Grassot, L., Dubuis, M., Faure, E., Couvidat, F., Caudeville, J., Bessagnet, B., Salizzoni, P., Leffondré, K., Gulliver, J., Severi, G., Mancini, F. R., & Fervers, B. (2022b). Long-term exposure to nitrogen dioxide air pollution and breast cancer risk: A nested case-control within the French E3N cohort study. *Environmental Pollution*, *317*, 120719. https://doi.org/10.1016/j.envpol.2022.120719.
14. Ssemugabo, C., Nalinya, S., Lubega, G. B., Ndejjo, R., & Musoke, D. (2020b). Health Risks in our Environment: Urban slum Youth' perspectives using photovoice in Kampala, Uganda. *Sustainability*, *13*(1), 248. https://doi.org/10.3390/su13010248.

Authors Introduction

Novi Ramadani

She is currently pursuing a Bachelor of Informatics with at the Faculty of Informatics, Telkom University Purwokerto, Indonesia. Her research interests include Artificial Intelligence, data analysis, and business analytics.

Andi Prademon Yunus, Ph. D

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.