Geographic and Risk Factor Analysis of Non-Communicable Cardiovascular Diseases in Central Java using Machine Learning

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

Abstract

Non-communicable diseases (NCDs), especially cardiovascular disease, are a major health problem in Indonesia with a global mortality rate of 17.9 million per year. In Central Java, hypertension cases will reach 8.5 million patients by 2023. This study uses the CART (Classification and Regression Tree) method and geographic mapping with Python to identify cardiovascular disease risk factors. Results showed alcohol stores as the dominant risk factor (54.3%), followed by sweet drinks (25.7%) and smokers (17.1%). The mapping identified the distribution of alcohol stores in 19 regions as a major factor in Central Java, Indonesia.

Keywords: Non-Communicable Diseases, Cardiovascular Diseases, Classification and Regression Tree, Central Java, Risk Factors, Geographic Distribution

1. Introduction

Non-communicable diseases (NCDs) are a major health problem in Indonesia, with the proportion of deaths reaching 76%. These non-communicable diseases mainly occur in low- and middle-income countries [1]. It is estimated that Indonesia experienced a total potential loss of 4.47 trillion US dollars from 2012 to 2030 due to non-communicable diseases. The high prevalence of NCDs can lead to increased demand for health services, more expensive treatment, and increased health expenditure, which in turn can reduce the budget available for investment in more productive activities [2].

Among the main types of NCDs, cardiovascular disease (CVD) is the leading cause of death in Indonesia with a proportion reaching 38%, followed by cancer at 12%, diabetes at 7%, chronic respiratory disease at 6%, and other NCDs at 13%. Cardiovascular disease is the leading cause of death worldwide, with 17.9 million deaths per year [3]. Cardiovascular diseases include various disorders of the heart and blood vessels, such as coronary heart disease, cerebrovascular disease, hypertension, peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis, and pulmonary embolism [4].

The main factors that cause cardiovascular disease include the consumption of unhealthy foods, lack of physical activity or calorie-burning exercise, alcohol consumption, smoking, and stress levels. Behavioral risk factors, such as unhealthy diet and lack of physical activity, along with smoking and alcohol consumption, are of major concern in cardiovascular disease prevention efforts [5]. The more risk factors a person has, the higher the likelihood of developing cardiovascular disease [6]. Therefore, a thorough analysis of these risk factors is crucial to identify influences and correlations associated with cardiovascular disease. Some regions in Indonesia, such as Central Java Province, record a high prevalence of cardiovascular disease, suggesting the importance of a deeper understanding of the risk factors that contribute to the disease.

Central Java Province is among the top five regions with the highest cardiovascular prevalence (hypertension) [2]. In 2023, the estimated number of people with hypertension in Central Java aged over 15 years reached 8,554,672 people or about 38.2 percent of the population of that age. This figure has increased compared to the previous year, indicating that hypertension is a serious health problem in this region [7].

This makes hypertension a major health problem that needs special attention. Therefore, the selection of Central Java as the location of this study is very appropriate to understand more about the geographical characteristics and risk factors that affect cardiovascular disease by utilizing machine learning or machine learning using the CART (Classification And Regression Tree) algorithm.

CART is used for decision-making related to classification and regression. In classification, CART produces a decision tree that maps observations into classes or categories. In regression, CART produces a decision tree that predicts the numerical or continuous value of the target variable [8]. This research will focus on regression or regression tree, the resulting regression tree is used to predict the numerical or continuous value of the target variable based on the relationship with the independent variables.

Therefore, this study aims to determine the CART method in identifying risk factors with the greatest influence on cardiovascular disease in Central Java. Risk factors with higher feature importance indicate that they have a greater contribution in causing cardiovascular disease. In addition, this research will also be accompanied by visualization of mapping the results of the analysis geographically, so as to know the geographical distribution of cardiovascular disease risk factors. This mapping will be created using the Python programming language and utilizing libraries such as GeoPandas to read and process geospatial data. The results of this research are expected to be useful not only

for health practitioners and policymakers in Central Java but also for similar research in other regions facing similar health challenges.

2. Methodology

This section describes the research methodology, including the data collection scheme, data preprocessing, and CART model design.

2.1. *Data Collection*

The cardiovascular disease dataset was obtained from the Central Java Health Office, which includes data from 2017 to 2023. However, in this study, only the 2023 data was used to ensure the relevance and accuracy of the analysis.

Meanwhile, risk factor data was collected through a scraping process from Google Maps using a Chrome extension called Instant Data Scraper. The scraping process was done manually for each region in Central Java, which consists of 29 districts and 6 cities.

Once the risk factor data and the number of cardiovascular disease patients have been combined, the dataset is now ready to be used for further analysis. However, before starting the analysis, another important step is to perform data preprocessing to ensure the data is clean, consistent, and ready to be optimally processed.

2.2. *Modeling*

In the modeling stage, the Decision Tree Regressor is used to predict the target variable, with logarithmic transformation on the target data to reduce the influence of outliers and ensure a more normal distribution. To improve model performance, hyperparameter tuning was performed using GridSearchCV with grid parameters including max_depth, min_samples_split, min_samples_leaf, and max_features. The tuning results showed the best parameters with max_depth 6, max_features 'sqrt', min_samples_leaf 2, and min_samples_split 2, which resulted in a model with lower prediction error based on neg_mean_squared_error evaluation. This combination of parameters allows the model to capture the data patterns better.

2.3. *Evaluation Model*

In this study, the model evaluation used is RMSE (Root Mean Squared Error) and R^2 (R-squared). RMSE measures the average error of the model prediction against the true value, the smaller the RMSE value or closer to 0, the better the model in prediction [9]. Meanwhile, R^2 measures how well the variation in data can be explained by the model, with values ranging from 0 to 1. The higher the R^2 value, the better the model is at explaining variations in the data [10]. The mathematical formulas of RMSE and R^2 are given in Eq. (1) and Eq. (2), respectively

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

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

2.4. *Feature Importance*

Feature Importance measures the contribution of each feature in predicting the target by calculating how much the feature reduces uncertainty in the decision tree. The value ranges from 0 to 1, with higher values indicating greater influence. Feature importance can help identify the risk factors that have the most influence on cardiovascular disease prevalence.

3. Results and Discussions

3.1. *Model Performance Evaluation*

The model evaluation results show an RMSE of 0.33, which indicates a very small prediction error and good predictive ability of the model on the available data. An RMSE close to 0 indicates better model performance, although these results are more relevant for data exploration than evaluation of new data. In addition, the R^2 value of 0.91 indicates that the model is able to explain 91% of the variation in the target data, indicating that the model is effective in identifying data patterns and providing accurate predictions, with only 9% of the variance unexplained.

3.2. *Feature Importance Result*

The feature importance results obtained from the analysis are presented in Table 1 and in the bar chart visualization below.

Table 1. Feature Importance

No.	Feature	Importance
1.	alcohol_store	0.2728
2.	sweet_drinks	0.2231
3.	smokers	0.1970
4.	transport	0.1269
5.	gym	0.0922
6.	fast_food	0.0631
7.	park	0.0217
8.	sports_center	0.0031
9.	tourist_spots	0.0000

The bar chart visualization in Fig. 1 illustrates the contribution of each feature in the model. In this graph, features are displayed on the Y-axis, and important values are shown on the X-axis.

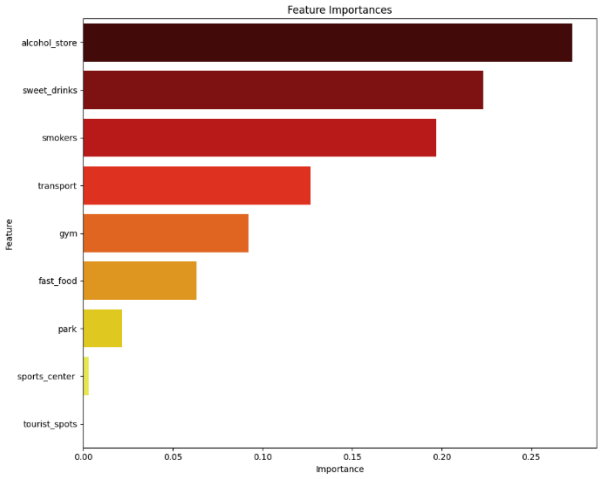

Fig.1. Feature Importance of Risk Factors on CART Model

1. Alcohol stores have the largest contribution (0.2728), indicating that the presence of alcohol stores plays a significant role in the increased risk of cardiovascular disease.
2. Sweet drinks (0.2231) and smokers (0.1970) also have large contributions to the prediction, with consumption of sweet drinks and smoking increasing the risk of cardiovascular disease.
3. Features with small contributions such as sports_center (0.0031) and tourist_spots (0.0000) indicate that sports facilities and tourist attractions have little effect on cardiovascular disease risk.

3.3. *Geographic Distribution*

The mapping of dominant risk factors in Central Java was conducted using matplotlib and geopandas, with Set3 colormaps to distinguish risk factor categories.

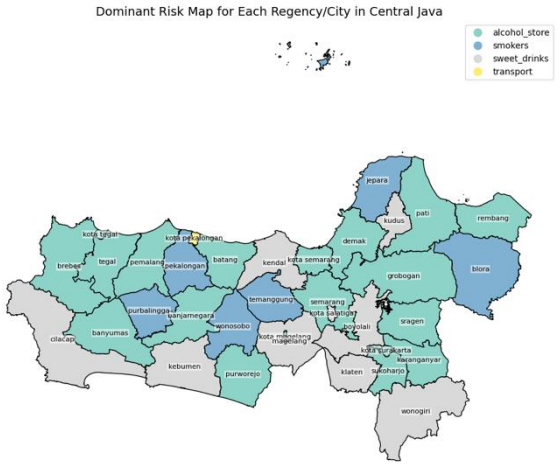

Fig.2. Mapping Results of Dominant Risk Factors in Central Java

The visualization results, as shown in Fig. 2, reveal that the alcohol store factor is the most dominant in 54.3% of the regions, followed by sweet_drinks (25.7%), smokers (17.1%), and transport (2.9%). These findings suggest that alcohol consumption is a major risk factor that needs more attention in cardiovascular disease prevention efforts in Central Java.

4. Conclusion

This study successfully identified and mapped the main risk factors for cardiovascular disease in Central Java using the CART method. Of the 9 factors analyzed, 4 factors were found to have a significant effect: alcohol_store (54.3%), sweet_drinks (25.7%), smokers (17.1%), and transport (2.9%), with alcohol_store as the dominant factor. The mapping shows that alcohol_store is distributed in 19 regions, sweet_drinks in 9 regions, smokers in 6 regions, and transportation in 1 region.

References

1. NCD Data Portal. “Noncommunicable Diseases Data Portal” [online]. Available : https://ncdportal.org/
2. Bloom, David & Chen, Simiao & McGovern, Mark & Prettner, Klaus & Candeias, Vanessa & Bernaert, Arnaud & Cristin, Stéphanie. Economics of Non-Communicable Diseases in Indonesia.
3. World Health Organization: WHO, “Noncommunicable diseases,” Dec. 23, 2024. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
4. Hassan, C. A. U., Iqbal, J., Irfan, R., Hussain, S., Algarni, A. D., Bukhari, S. S. H., Alturki, N., & Ullah, S. S. (2022). Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. Sensors (Basel, Switzerland), 22(19), 7227. https://doi.org/10.3390/s22197227
5. World Health Organization: WHO, “Cardiovascular diseases (CVDs),” Jun. 11, 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
6. H. E. Bays et al., “Ten things to know about ten cardiovascular disease risk factors,” American Journal of Preventive Cardiology, vol. 5, p. 100149, Jan. 2021, doi: 10.1016/j.ajpc.2021.100149.
7. D. N. Khasanah, “The Risk Factors Of Hypertension In Indonesia (Data Study Of Indonesian Family Life Survey 5),” Journal of Public Health Research and Community Health Development, vol. 5, no. 2, p. 80, Feb. 2022, doi: 10.20473/jphrecode.v5i2.27923.
8. W. Dong et al., “Risk factors and geographic disparities in premature cardiovascular mortality in US counties: a machine learning approach,” Scientific Reports, vol. 13, no. 1, Feb. 2023, doi: 10.1038/s41598-023-30188-9.
9. W. A. C. Castañeda and P. B. Filho, “Improvement of an Edge-IoT architecture driven by artificial intelligence for Smart-Health chronic disease management,” Sensors, vol. 24, no. 24, p. 7965, Dec. 2024, doi: 10.3390/s24247965.
10. S. Panda, B. Purkayastha, D. Das, M. Chakraborty, and S. K. Biswas, “Health insurance cost prediction using regression models,” 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), pp. 168–173, May 2022, doi: 10.1109/com-it-con54601.2022.9850653.

Authors Introduction

Ms. Nurhasanah

She is an undergraduate student of Informatics Engineering at Telkom University, Indonesia with a GPA of 3.96. She has expertise in data analysis, human resource development, social media specialist, and copywriting.

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.