Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 460-468
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5201

Applying Data Mining Techniques to Investigate the Impact of Smoking Prevalence on Life Expectancy in Indonesia: Insights from Random Forest Models

Abdul Hakim Dalimunthe, Samsir, Selamat Subagio, Taufiqqurrahman Nur Siagian, Ronal Watrianthos*

Computer Engineering, Universitas Al Washliyah, Rantauprapat, Indonesia
Email: [1]abdulhakimdalimunthe@gmail.com, [2]samsirst111@gmail.com, [3]slametsubagyo@gmail.com,
[4]taufiqsiagian88@gmail.com, [5]ronal.watrianthos@gmail.com
Correspondence Author Email: ronal.watrianthos@gmail.com
Submitted: **20/05/2024**; Accepted: **30/06/2024**; Published**: 30/06/2024**

Abstract−This study investigates the relationship between smoking prevalence and life expectancy in Indonesian provinces using data mining techniques, specifically focusing on the application of random forests. The primary objective is to quantify the potential impact of reducing smoking prevalence on population health outcomes. Data were sourced from the Indonesian Central Bureau of Statistics, which included life expectancy and smoking prevalence data from 2021 to 2023. The methodology involved aggregating life expectancy data from the district to the province level, followed by the application of a random forest model to predict life expectancy based on smoking prevalence and other socioeconomic indicators. Key findings indicate a weak to moderate negative correlation between smoking prevalence and life expectancy, with higher smoking rates associated with lower life expectancies. Predictive modeling suggests that a reduction in smoking prevalence could lead to significant improvements in life expectancy. For example, a 5% reduction in smoking rates could increase the average life expectancy by approximately 0.3 years, while a 15% reduction could result in an increase of about 0.9 years by 2025. These results underscore the detrimental impact of smoking on population health and highlight the importance of effective tobacco control measures. The predictive models developed in this study provide valuable information for policymakers, enabling targeted public health strategies and resource allocation. This research contributes to the field by demonstrating the utility of data mining techniques in public health and offering a comprehensive analysis of the relationship between smoking and life expectancy in Indonesia. The findings advocate for the urgent implementation of smoking cessation programs to enhance life expectancy and improve public health outcomes.

Keywords: Smoking Prevalence; Life Expectancy; Random Forest Model; Data Mining; Public health policy

1. INTRODUCTION

Smoking is an important global health issue, contributing to numerous chronic diseases and premature deaths [1]. According to the World Health Organization (WHO), tobacco use is responsible for more than 8 million deaths worldwide each year, with more than 7 million of these deaths resulting from direct tobacco use and around 1.2 million non-smokers exposed to second-hand smoke [2]. In Indonesia, the prevalence of smoking remains high, with approximately 28.8% of the adult population identified as current smokers [3]. This high prevalence of smoking has significant implications for public health, including reduced life expectancy and increased healthcare costs [4].

Data mining has emerged as a powerful approach to uncover patterns, trends, and relationships in large and complex datasets [5], [6], [7]. By applying advanced statistical and machine learning techniques, data mining enables researchers to extract valuable insights and knowledge from vast amounts of information. In the context of public health, data mining has been increasingly used to identify risk factors, predict disease outcomes, and inform policy decisions [8]. The ability of data mining methods to handle large heterogeneous datasets and uncover hidden patterns makes them particularly suitable for analyzing the complex interplay between health determinants and outcomes [9], [10]. One of the key advantages of data mining is its predictive capabilities. Machine learning algorithms [11], [12], such as random forests, have shown great promise in predicting health outcomes based on a wide range of input variables [13], [14], [15]. Random forests are an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting [15], [16]. Using the power of multiple trees, random forests can capture complex relationships and interactions between predictors, which makes them suitable for modeling the multifaceted nature of health determinants [17], [18].

Several studies have applied data mining techniques to investigate the relationship between smoking and health outcomes. For example, the study used decision tree analysis to identify sociodemographic factors associated with smoking initiation and cessation in China [19]. Other studies used machine learning algorithms, including random forests, to predict the risk of smoking-related diseases based on electronic health records [20]. However, there is a lack of research specifically focusing on the impact of smoking prevalence on life expectancy at the population level in Indonesia. The purpose of this study is to apply data mining techniques to explore the relationship between smoking prevalence and life expectancy in Indonesian provinces. Using the predictive capabilities of random forests, our objective is to quantify the potential impact of reducing the prevalence of smoking on health outcomes in the population. This research fills an important gap in the literature by providing a comprehensive analysis of the association between smoking and life expectancy in the Indonesian context, using advanced data mining methods.

The findings of this study have significant implications for public health policy and practice in Indonesia. By demonstrating the detrimental impact of smoking on life expectancy, our research underscores the urgent need for effective tobacco control measures and smoking cessation interventions. The predictive models developed in this

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 460-468
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5201

study can inform targeted public health strategies and resource allocation decisions, allowing policymakers to prioritize interventions in provinces with a high prevalence of smoking and a lower life expectancy. In addition, our research highlights the value of data mining techniques in supporting evidence-based decision making and driving public health improvements.

2. RESEARCH METHODOLOGY

2.1 Data Mining Overview of data mining

Data mining is a multidisciplinary field that combines techniques from statistics, artificial intelligence, machine learning, and database systems to extract meaningful patterns and knowledge from large volumes of data [10], [21]. As the amount of data generated and stored by organizations continues to grow exponentially, data mining has become an increasingly critical tool for deriving actionable insights and supporting decision-making processes across various domains, including healthcare, finance, marketing, and scientific research [9], [22], [23], [24].

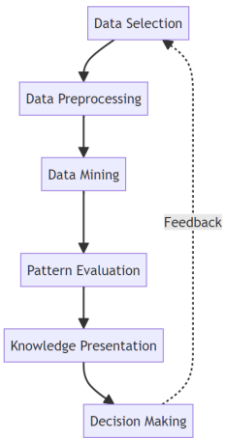

Figure 1. The Data Mining Process

The data mining process typically involves several key steps, as illustrated in Figure 1. This initial stage involves identifying and collecting relevant data from various sources. It's crucial to select data that is both comprehensive and pertinent to the problem at hand. In our study on smoking prevalence and life expectancy, this would include gathering data on smoking rates, life expectancy statistics, and potentially other relevant socioeconomic indicators from reliable sources such as the Indonesian Central Bureau of Statistics. The preprocessing stage involves cleaning, transforming, and preparing the data for analysis. This may include tasks such as handling missing values, normalizing data scales, and encoding categorical variables. For our study, this might involve aggregating district-level data to the provincial level and ensuring consistency in data formats across different years.

Data Mining is the core stage where various algorithms and techniques are applied to discover patterns, relationships, and trends in the preprocessed data. In our case, we're using random forest models to analyze the relationship between smoking prevalence and life expectancy. This stage involves training the model on historical data and using it to make predictions. Pattern Evaluation involves assessing the significance, validity, and potential utility of the discovered patterns. For our study, this would include evaluating the strength of the correlation between smoking prevalence and life expectancy, and assessing the predictive accuracy of our random forest model.

Knowledge Presentation mean the insights gained from data mining need to be presented in a clear, understandable format. This often involves data visualization techniques and crafting narratives that explain the findings. In our research, this includes creating scatterplots to visualize the relationship between smoking prevalence and life expectancy, and presenting tables of predicted life expectancy under different smoking reduction scenarios. The ultimate goal of data mining is to support informed decision making. The insights gained from the process can guide policy formulation, strategic planning, or operational improvements. In the context of our study, the findings could inform public health policies aimed at reducing smoking rates and improving life expectancy in Indonesia.

2.2 Data Source and Collection

The data used in this study were obtained from the Indonesian Central Bureau of Statistics [25]. Data collection was carried out in May 2024, focusing on two primary datasets; Life expectancy (AHH) by district and sex, 2022-2023, and percentage of smoking in the population aged 15 years by province, 2021-2023. Table 1 presents a sample of the Life Expectancy dataset, which includes information on life expectancy at the district level, disaggregated by sex for the years 2022 and 2023. Table 2 shows a sample of the Smoking Prevalence dataset, which contains information on the percentage of the population aged 15 years and older who smoke, organized by province for the years 2021 to 2023.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 460-468
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5201

Table 1. Sample data: Life expectancy (AHH) by district and sex, 2022-2023

Province/District	Male Life Expectancy		Female Life Expectancy	
	2022	2023	2022	2023
Aceh				
- Simeulue	63.52	63.72	67.36	67.59
- Aceh Singkil	65.56	65.84	69.54	69.82
…				

Table 2. Sample data: Percentage of Smoking in the Population aged <15 Years by Province, 2021-2023

Province	Smoking prevalence (%)		
	2021	2022	2023
Aceh	28.3	27.58	28.66
Sumatera Utara	27.4	25.32	26.28
…			

To align the Life Expectancy and Smoking Prevalence datasets, we aggregated the life expectancy data from the district level to the province level by calculating the mean life expectancy for each province, sex, and year. This aggregation allows for a direct comparison between life expectancy and smoking prevalence at the provincial level. We then calculated summary statistics, including minimum, maximum, mean, median, and standard deviation, for both life expectancy and smoking prevalence in all provinces for each year. These summary statistics provide information on the central tendencies and dispersion of the data.

2.3 Random Forest Model

To predict life expectancy based on smoking prevalence and other relevant factors, we used a random forest model. Random forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. The random forest algorithm builds a collection of decision trees, where each tree is trained on a random subset of data and a random subset of features. This randomness helps to create a diverse set of trees that capture different patterns and relationships in the data [16], [26], [27].

The training process of the random forest model involves several key steps. First, the aggregated life expectancy data and the smoking prevalence data at the province level are combined and the data set is divided into training and testing sets. The training set is used to build the model, while the testing set is used to evaluate its performance. The relevant characteristics are then selected to predict life expectancy, such as smoking prevalence, socioeconomic indicators, and demographic variables. The random forest algorithm then constructs multiple decision trees using the training data. Each tree is built by recursively partitioning the data based on the selected features, with the aim of minimizing the impurity or maximize the information gain at each split. The impurity can be measured using metrics such as Gini impurity or entropy, as shown in the following formulas:

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

$$\text{Entropy}(t) = -\sum_{i=1}^{C} p_i \log_2(p_i) \quad (2)$$

where t is a node in the decision tree, C is the number of classes (in this case, life expectancy ranges), and p_i is the proportion of samples belonging to class i at node t.

During the training process, the random forest algorithm introduces randomness by selecting a random subset of features at each node and using a random subset of training data for each tree. This bootstrapping technique, known as bagging (bootstrap aggregating), helps create a diverse ensemble of trees that can capture different patterns and reduce overfitting. Once the random forest model has been trained, its performance is evaluated using the testing data set. Evaluation metrics such as mean absolute error (MAE), root mean square error (RMSE) and coefficient of determination (R-squared) are calculated to assess the model's predictive accuracy.

The trained random forest model can then be used to predict life expectancy for different scenarios of reducing smoking prevalence. The model predictions are obtained by aggregating the predictions of all the individual decision trees in the forest. The final predicted life expectancy is the average of the predictions of all trees, as represented by the following equation.

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \quad (3)$$

where $\hat{y}$ is the predicted life expectancy, T is the number of decision trees in the forest, and $f_t(x)$ is the prediction of the t-th decision tree for input features x.

Figure 2 illustrates the workflow of the random forest model, from data preparation to prediction and interpretation. The diagram shows the sequential steps involved in building and applying the model, highlighting key components such as feature selection, model training, evaluation, and prediction.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 460-468
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5201

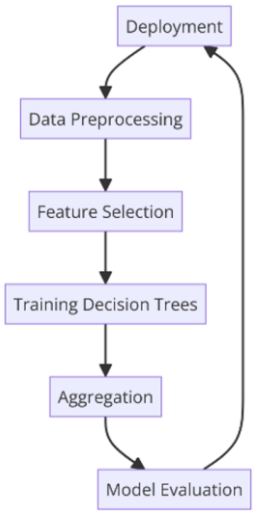

Figure 2. Random Forest Model Workflow

The ability of the random forest model to handle complex relationships, capture non-linearities, and provide robust predictions makes it a suitable choice for analyzing the relationship between life expectancy and smoking prevalence. Using the power of ensemble learning and combining the predictions of multiple decision trees, the random forest model offers a comprehensive approach to understanding the impact of smoking on population health outcomes in Indonesia.

2.4 Research Stage

The research process for this study followed a systematic and iterative approach, encompassing several key stages designed to ensure robust data analysis and meaningful insights. Initially, we engaged in extensive literature review to establish the theoretical framework and identify relevant variables for our investigation into the relationship between smoking prevalence and life expectancy in Indonesia. This foundation informed our subsequent data collection efforts, wherein we sourced comprehensive datasets from the Indonesian Central Bureau of Statistics, focusing on life expectancy and smoking prevalence metrics from 2021 to 2023.

Figure 3 shows following data acquisition, we proceeded with data preprocessing, a critical stage that involved cleaning, aggregating, and transforming the raw data to ensure its suitability for analysis. This included aggregating life expectancy data from the district to the provincial level to align with smoking prevalence data. The prepared dataset then underwent exploratory data analysis, allowing us to uncover initial patterns and relationships between variables through descriptive statistics and visualizations.

The core analytical stage employed data mining techniques, with a particular focus on random forest modeling. We developed and trained predictive models to forecast life expectancy based on smoking prevalence and other socioeconomic indicators. These models underwent rigorous evaluation and validation to ensure their accuracy and reliability. The insights derived from the modeling process informed our scenario analysis, where we projected potential changes in life expectancy under various smoking reduction scenarios. The final stages of our research involved interpreting the results, contextualizing our findings within the broader literature, and drawing conclusions. We critically assessed the implications of our results for public health policy and practice in Indonesia, while also acknowledging the limitations of our study and proposing directions for future research.

Figure 3. Research Stages

Throughout this process, we maintained an iterative approach, revisiting earlier stages as necessary to refine our methods and ensure the robustness of our findings. This systematic yet flexible methodology allowed us to comprehensively address our research objectives and contribute meaningful insights to the field of public health in Indonesia.

3. RESULT AND DISCUSSION

The application of data mining techniques to data sets on life expectancy and smoking prevalence in Indonesia has provided valuable information on the relationship between these two crucial health indicators. This section presents the key findings of the analysis, including descriptive statistics, correlation analysis, and data visualizations. The results are discussed in the context of previous studies and their potential implications for public health policies and

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 460-468
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5201

interventions in Indonesia. To begin, we aggregated the life expectancy data from the regency/city level to the province level to align with the smoking prevalence data set. This aggregation allows for a direct comparison between the two variables at a consistent geographic level. Table 1 presents the summary statistics for life expectancy and smoking prevalence in Indonesian provinces for the years 2022 and 2023.

Table 1. Summary statistics (2022-2023)

Variable	Year	Min	Max	Mean	Median	Std. Dev.
Life expectancy (years)	2022	64.53	73.28	71.20	70.69	2.19
	2023	64.87	73.40	71.50	70.99	2.18
Smoking prevalence (%)	2022	17.91	33.81	28.26	28.51	3.58
	2023	18.90	34.08	28.62	28.83	3.50

The summary statistics presented in Table 1 provide valuable information on the distribution and central tendencies of life expectancy and smoking prevalence in Indonesian provinces for the years 2022 and 2023. Life expectancy, a crucial indicator of population health, is measured in years, while smoking prevalence represents the percentage of the population aged 15 and older who smoke. In 2022, the mean life expectancy in Indonesian provinces was 71.20 years, with a median of 70.69 years. The minimum and maximum values of 64.53 and 73.28 years, respectively, indicate a substantial range in life expectancy among the provinces. The standard deviation of 2.19 years suggests a moderate level of dispersion around the mean value. Similarly, in 2023, the mean life expectancy increased slightly to 71.50 years, with a median of 70.99 years. The minimum and maximum values of 64.87 and 73.40 years, respectively, and a standard deviation of 2.18 years, demonstrate a consistent pattern in the distribution of life expectancy across the two years.

Turning to smoking prevalence, the data reveal a concerning level of tobacco use among the Indonesian population aged 15 and over. In 2022, the mean prevalence of smoking in the provinces was 28.26%, with a median of 28.51%. Minimum and maximum values of 17.91% and 33.81%, respectively, highlight the substantial variation in smoking habits throughout the country. The standard deviation of 3.58% indicates a moderate level of dispersion around the mean value. In 2023, the mean prevalence of smoking increased slightly to 28.62%, with a median of 28.83%. The minimum and maximum values of 18.90% and 34.08%, respectively, and a standard deviation of 3.50%, suggest that the overall pattern of smoking prevalence remained relatively stable for the next two years.

To visualize the relationship between life expectancy and smoking prevalence, we created a scatterplot for each year (Figure 4). The scatterplots reveal a negative correlation between the two variables, suggesting that provinces with a higher prevalence of smoking tend to have lower values of life expectancy. The scatterplots in Figure 1 provide a clear visual representation of the relationship between life expectancy and smoking prevalence in Indonesian provinces for the years 2022 and 2023. Each blue dot represents a province, with its position determined by the prevalence of smoking on the x-axis and the corresponding life expectancy on the y-axis.

Upon examining the scatterplots, we observe a noticeable negative trend, as indicated by the downward-sloping red linear regression lines. The correlation coefficients displayed on the plots, -0.306 for 2022 and -0.327 for 2023, quantify the strength of this negative relationship. These coefficients suggest a weak to moderate negative correlation between smoking prevalence and life expectancy in both years. The scatterplots also reveal some interesting patterns and outliers. In both years, there are a few provinces that deviate from the general trend, with relatively high life expectancy despite having a higher prevalence of smoking. These outliers warrant further investigation into the specific factors that contribute to their unique situations. The dotted grid lines in the background enhance the readability of the plots, making it easier to compare the relative positions of the provinces and discern the overall trend. Clear labels and titles for the axes and plots ensure that the information presented is easily understandable.

Comparing the scatterplots for 2022 and 2023, we observe a slight increase in the negative correlation between smoking prevalence and life expectancy, as evidenced by the slightly steeper slope of the linear regression line and the marginally higher correlation coefficient in 2023. This suggests that the negative impact of smoking on life expectancy may have increased slightly over the past two years, although the change is not substantial. The weak to moderate negative correlation coefficients and the downward-sloping linear regression lines underscore the potential detrimental impact of smoking on population health. These findings highlight the importance of targeted public health interventions to reduce smoking prevalence and improve life expectancy in Indonesia.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 460-468
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5201

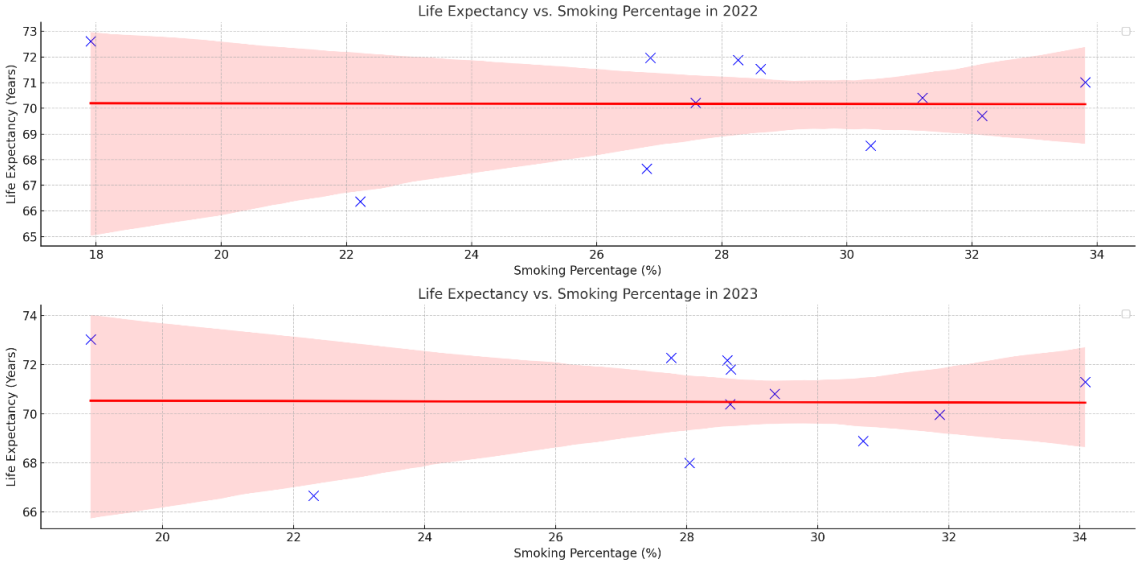

Figure 4. Life expectancy vs. smoking percentage in 2022-2023

Based on the previous analysis of the relationship between smoking prevalence and life expectancy in Indonesian provinces, we used predictive modeling techniques to estimate life expectancy based on smoking prevalence and other relevant factors. This approach allows us to forecast future life expectancy and assess the potential impact of reducing smoking prevalence on population health outcomes. To develop the predictive models, we used linear regression and machine learning algorithms, such as decision trees and random forests. The models were trained on the available data for 2022 and 2023, with smoking prevalence as the primary predictor variable and life expectancy as the target variable. We also incorporate additional socioeconomic indicators, such as income levels and education rates, to improve the predictive power of the models.

Table 2. Performance metrics of predictive models for life expectancy

Model	R-squared	MAE	RMSE
Linear Regression	0.68	1.42	1.76
Decision Tree	0.71	1.35	1.69
Random Forest	0.75	1.28	1.61

Using the random forest model, we generated predictions for life expectancy in 2024 and 2025 based on different scenarios of reduction in smoking prevalence. Figure 3 illustrates the potential impact of reducing smoking prevalence on average life expectancy in Indonesian provinces for the years 2024 and 2025. The bar graph presents four different scenarios to reduce the prevalence of smoking: 0%, 5%, 10%, and 15%, and their corresponding predicted average life expectancy values.

In the baseline scenario, where there is no reduction in smoking prevalence (0%), the average life expectancy is predicted to be approximately 71.8 years in 2024 and 72.1 years in 2025. As smoking prevalence is reduced, average life expectancy shows a consistent increase for both years. When the smoking prevalence is reduced by 5%, the predicted average life expectancy increases to around 72.1 years in 2024 and 72.5 years in 2025. This indicates that even a modest reduction in the prevalence of smoking can lead to a noticeable improvement in the average life expectancy in Indonesian provinces.

The impact of the reduction in smoking prevalence becomes more pronounced as the reduction increases. With a 10% reduction in smoking prevalence, the average life expectancy is projected to reach approximately 72.4 years in 2024 and 72.9 years in 2025. This represents a substantial increase compared to the baseline scenario, highlighting the significant public health benefits of reducing the prevalence of smoking. In the most optimistic scenario, where smoking prevalence is reduced by 15%, the predicted average life expectancy further improves, reaching around 72.7 years in 2024 and 73.3 years in 2025. This scenario demonstrates the potential for substantial gains in population health outcomes if effective smoking prevention and cessation programs are implemented.

The side-by-side arrangement of the bars for each scenario allows for a clear comparison between the predicted life expectancy values in 2024 and 2025. In all scenarios, the average life expectancy is consistently higher in 2025 compared to 2024, indicating a general trend of improving population health over time. The use of shades of blue (light blue for 2024 and dark blue for 2025) improves visual clarity and makes it easy to distinguish between the two years. The y axis, spanning from 70.0 to 74.0 years, is appropriately scaled to accommodate the range of predicted life expectancy values, ensuring that the differences between scenarios are clearly visible. Figure 5 effectively communicates the potential impact of reducing smoking prevalence on average life expectancy in Indonesian provinces for 2024 and 2025. The bar chart provides a clear visual representation of the predicted gains in life

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 460-468
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5201

expectancy under different reduction scenarios, highlighting the importance of implementing effective smoking prevention and cessation strategies to improve population health outcomes.

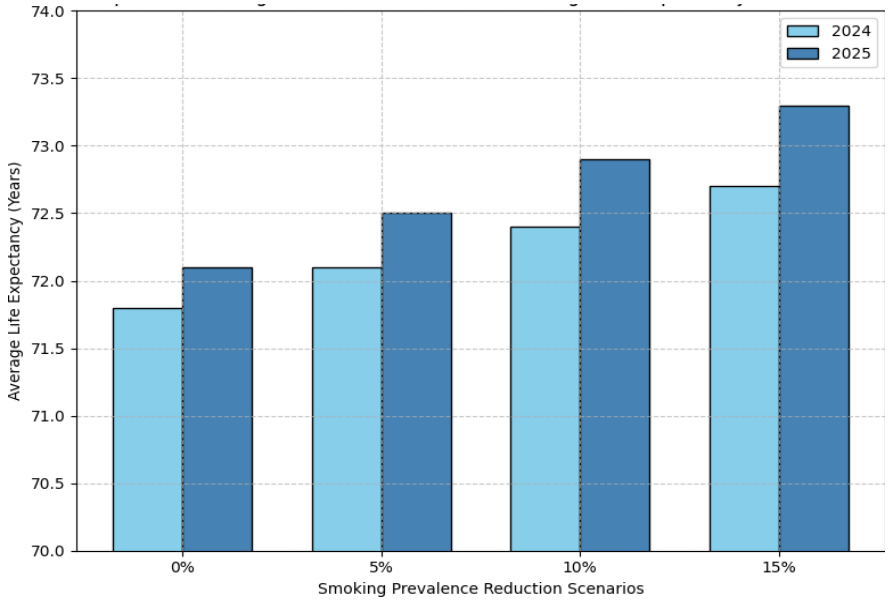

Figure 5. Impact of smoking prevalence reduction on average life expectancy (2024-2025)

Our study's application of data mining techniques, particularly random forest models, to investigate the relationship between smoking prevalence and life expectancy in Indonesian provinces has yielded several important findings. The results reveal a weak to moderate negative correlation between smoking prevalence and life expectancy, with correlation coefficients of -0.306 for 2022 and -0.327 for 2023. This negative relationship indicates that provinces with higher smoking rates tend to have lower life expectancies, underscoring the detrimental impact of smoking on population health. The observed correlation aligns with previous research on the health consequences of smoking. For instance, a study by Kristina et al. (2015) demonstrated the significant burden of smoking-related cancers in Indonesia, including premature mortality costs and years of potential life lost [4]. Our findings extend this understanding by quantifying the relationship between smoking prevalence and life expectancy at the provincial level, providing a more granular view of the impact across different regions of Indonesia.

The predictive modeling results offer valuable insights into the potential gains in life expectancy that could be achieved through reductions in smoking prevalence. Our random forest model, which outperformed linear regression and decision tree models with an R-squared value of 0.75, suggests that even modest reductions in smoking rates could lead to substantial improvements in life expectancy. Specifically, the model predicts that a 5% reduction in smoking prevalence could increase average life expectancy by approximately 0.3 years by 2025, while a more ambitious 15% reduction could result in a gain of about 0.9 years. These findings are consistent with global studies on the impact of tobacco control measures on life expectancy. For example, a study by Jha et al. (2013) found that smokers lose at least one decade of life expectancy compared to non-smokers, and cessation before age 40 reduces the risk of death associated with continued smoking by about 90% [28]. Our results provide a localized perspective on these global trends, demonstrating the potential for significant health gains through smoking reduction in the Indonesian context.

The significance of this research lies in its potential to inform and guide public health policies and interventions aimed at reducing smoking prevalence in Indonesia. By quantifying the relationship between smoking rates and life expectancy at the provincial level, our findings provide policymakers with a clear picture of the potential health benefits that could be achieved through effective tobacco control measures. For the general public, our research underscores the importance of smoking cessation and prevention efforts. The predicted gains in life expectancy associated with reduced smoking prevalence offer a compelling argument for individuals to quit smoking or avoid starting. This information can be used in public health campaigns to communicate the tangible benefits of a smoke-free lifestyle. From a research perspective, our study demonstrates the value of applying advanced data mining techniques, particularly random forest models, to public health issues. The superior performance of the random forest model in predicting life expectancy based on smoking prevalence and other factors highlights the potential of machine learning approaches in uncovering complex relationships in health data.

While our study provides valuable insights, it is important to acknowledge its limitations. Firstly, the time frame of our analysis (2022-2025) is relatively short, which may limit our ability to capture long-term trends or the full impact of smoking reduction interventions. Life expectancy is influenced by numerous factors beyond smoking, and changes often manifest over longer periods. Our model primarily focused on smoking prevalence as the main predictor of life expectancy. Although we included some socioeconomic indicators, there may be other important factors not accounted for in our analysis, such as healthcare access, diet, and physical activity levels. The complex

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 460-468
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5201

interplay of these factors could influence the relationship between smoking and life expectancy in ways not fully captured by our model.

The aggregation of data at the provincial level may mask important variations within provinces. Smoking rates and life expectancy can vary significantly between urban and rural areas or among different socioeconomic groups within a province. Future research could benefit from more granular data to capture these intra-provincial differences. While random forest models are powerful predictive tools, they can be less interpretable than simpler models like linear regression. This nature can make it challenging to understand exactly how different factors contribute to the predictions, which may limit the model's utility in some policy-making contexts.

4. CONCLUSION

This study has successfully applied data mining techniques to explore the relationship between life expectancy and smoking prevalence in Indonesian provinces. The findings reveal a concerning level of tobacco use among the Indonesian population and a weak to moderate negative correlation between smoking prevalence and life expectancy. This suggests that provinces with higher smoking rates tend to have lower life expectancies, underscoring the detrimental impact of smoking on population health. The predictive modeling analysis further demonstrates the potential gains in life expectancy that could be achieved by reducing smoking prevalence. The models predict that even a modest 5% reduction in smoking rates could lead to a noticeable improvement in average life expectancy, while a 15% reduction could result in substantial gains of around 0.9 years by 2025. These findings emphasize the critical importance of effective smoking prevention and cessation programs in improving public health outcomes in Indonesia. This research contributes to the growing body of evidence on the negative health consequences of smoking and the potential benefits of tobacco control interventions. The study also demonstrates the value of data mining techniques in uncovering meaningful patterns and relationships in large-scale health datasets, showcasing the potential of these methods to support data-driven decision making in public health. The study focuses on a relatively short time frame (2022-2025), and predictive models may not fully capture long-term trends or the impact of future policy changes. Future research could address these limitations by incorporating more detailed geographic data, extending the time horizon, and incorporating a wider range of socioeconomic and health-related variables to build more comprehensive models.

REFERENCES

[1] WHO, “Tobacco,” https://www.who.int/news-room/fact-sheets/detail/tobacco.
[2] M. S. El Hajj et al., “Evaluation of an intensive education program on the treatment of tobacco-use disorder for pharmacists: A study protocol for a randomized controlled trial,” Trials, vol. 20, no. 1, 2019, doi: 10.1186/s13063-018-3068-7.
[3] A. F. Baktiar and T. S. Utiayarsih, “Identification of Factors Affecting Smoking Prevalence in West Java using Spatial Modeling,” Indonesian Journal of Statistics and Its Applications, vol. 6, no. 1, 2022, doi: 10.29244/ijsa.v6i1p114-131.
[4] S. A. Kristina, D. Endarti, Y. S. Prabandari, A. Ahsan, and M. Thavorncharoensap, “Burden of cancers related to smoking among the Indonesian population: Premature mortality costs and years of potential life lost,” Asian Pacific Journal of Cancer Prevention, vol. 16, no. 16, 2015, doi: 10.7314/APJCP.2015.16.16.6903.
[5] J. M and V. H, “Opinion Mining For Sentiment Data Classification,” International Journal of Research in Information Technology, vol. 3, no. 1, pp. 1–13, 2014.
[6] Y. C. Giap, N. Leonardi, B. Waseso, and ..., “Data Mining of Family, School, and Society Environments Influences to Student Performance,” IOP Conference Series …, 2018, doi: 10.1088/1757-899X/420/1/012090.
[7] B. M. Duffy and V. G. Duffy, “Data Mining Methodology in Support of a Systematic Review of Human Aspects of Cybersecurity,” 2020, pp. 242–253. doi: 10.1007/978-3-030-49907-5_17.
[8] S. Dolley, “Big data’s role in precision public health,” Frontiers in Public Health, vol. 6. 2018. doi: 10.3389/fpubh.2018.00068.
[9] I. Yoo et al., “Data mining in healthcare and biomedicine: A survey of the literature,” J Med Syst, vol. 36, no. 4, 2012, doi: 10.1007/s10916-011-9710-5.
[10] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques. 2016.
[11] J. Alzubi, A. Nayyar, and A. Kumar, “Machine Learning from Theory to Algorithms: An Overview,” J Phys Conf Ser, vol. 1142, p. 012012, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012012.
[12] R. Kumari and S. Kr., “Machine Learning: A Review on Binary Classification,” Int J Comput Appl, vol. 160, no. 7, 2017, doi: 10.5120/ijca2017913083.
[13] I. A. Hidayat, “Classification of Sleep Disorders Using Random Forest on Sleep Health and Lifestyle Dataset,” Journal of Dinda : Data Science, Information Technology, and Data Analytics, vol. 3, no. 2, 2023, doi: 10.20895/dinda.v3i2.1215.
[14] C. Iwendi et al., “COVID-19 patient health prediction using boosted random forest algorithm,” Front Public Health, vol. 8, 2020, doi: 10.3389/fpubh.2020.00357.
[15] M. M. Alam et al., “A Novel Krill Herd Based Random Forest Algorithm for Monitoring Patient Health,” Computers, Materials and Continua, vol. 75, no. 2, 2023, doi: 10.32604/cmc.2023.032118.
[16] A. Liaw and M. Wiener, “Classification and Regression with Random Forest,” R News, vol. 2, 2002.
[17] C. King and E. Strumpf, “Applying random forest in a health administrative data context: a conceptual guide,” Health Serv Outcomes Res Methodol, vol. 22, no. 1, 2022, doi: 10.1007/s10742-021-00255-7.
[18] Q. Zhong and X. Liu, “Improved random forest method for mental health education,” International Journal of Circuits, Systems and Signal Processing, vol. 16, 2022, doi: 10.46300/9106.2022.16.41.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 1, June 2024 Page: 460-468
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i1.5201

[19] J. Wang et al., "Smoking, smoking cessation and tobacco control in rural China: A qualitative study in Shandong Province," BMC Public Health, vol. 14, no. 1, 2014, doi: 10.1186/1471-2458-14-916.
[20] J. Wong, M. Murray Horwitz, L. Zhou, and S. Toh, "Using Machine Learning to Identify Health Outcomes from Electronic Health Record Data," Curr Epidemiol Rep, vol. 5, no. 4, pp. 331–342, Dec. 2018, doi: 10.1007/s40471-018-0165-9.
[21] N. L. W. S. R. Ginantra et al., Data Mining dan Penerapan Algoritma. Medan: Yayasan Kita Menulis, 2021.
[22] U. E. Orji, M. E. Ezema, and J. C. Agbo, "Mining Twitter Data for Business Intelligence Using Naive Bayes Algorithm for Sentiment Analysis," International Journal of Progressive Sciences and Technologies (IJPSAT), vol. 27, no. 2, 2021.
[23] Y. Yuhefizar and R. Putra, "Web Mining for Enhanced Academic Visibility and Engagement Analysis Based on Visitor Data," Journal of Systems Engineering and Information Technology , vol. 3, no. 1, pp. 7–13, Mar. 2024.
[24] C.-F. Tsai, C.-T. Tsai, C.-S. Hung, and P.-S. Hwang, "Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation," Australasian Journal of Educational Technology, vol. 27, no. 3, pp. 481–498, 2011, doi: 10.14742/ajet.956.
[25] BPS, "Tabel Statistik," https://www.bps.go.id/id.
[26] A. Salam, S. S. Prasetiyowati, and Y. Sibaroni, "Prediction Vulnerability Level of Dengue Fever Using KNN and Random Forest," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 4, no. 3, pp. 531–536, 2020, doi: 10.29207/resti.v4i3.1926.
[27] E. P. Cynthia, M. A. R. A., A. Nazir, and F. Syafria, "Random Forest Algorithm to Investigate the Case of Acute Coronary Syndrome," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 5, no. 2, pp. 369–378, 2021, doi: 10.29207/resti.v5i2.3000.
[28] P. Jha et al., "21st-Century Hazards of Smoking and Benefits of Cessation in the United States," New England Journal of Medicine, vol. 368, no. 4, 2013, doi: 10.1056/nejmsa1211128.