JoCoSiR
Journal of Computer Science Research
ISSN 2986-2337 (Online)

Predicting Public Health Risks Based on Lifestyle Factors Using the Support Vector Machine

Andri Ismail Sitepu[1] , *Muhammad Iqbal*[2]

[1,2] *Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Kota Medan, Indonesia*

email: [1] *adrismstp@gmail.com*, [2]*wakbalpb@yahoo.co.id*

ARTICLE INFO

Keywords:
support vector machine,
health risk prediction,
lifestyle factors,
machine learning,
preventive healthcare

IEEE style in citing this article:
A.I Sitepu and M. Iqbal, " Predicting Public Health Risks Based on Lifestyle Factors Using the Support Vector Machine," *JoCoSiR: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 3, no. 3, pp. 62-66, 2025.

ABSTRACT

Public health risks are often influenced by multiple lifestyle factors, such as age, diet, exercise, smoking, and alcohol consumption. This study aims to develop a predictive model for assessing individual health risks using the Support Vector Machine (SVM) algorithm. The dataset used consists of lifestyle attributes, including age, weight, height, exercise frequency, sleep duration, sugar intake, smoking habits, alcohol consumption, marital status, profession, and body mass index (BMI). The data were preprocessed through normalization and label encoding, followed by training and testing using a 70:30 data split. The SVM model employed the Radial Basis Function (RBF) kernel to capture non-linear relationships between variables. Experimental results show that the proposed SVM model achieved an accuracy of approximately 89%, demonstrating strong predictive capability. The confusion matrix analysis revealed that the model effectively distinguishes between high and low health risk categories, while the PCA visualization confirmed clear clustering of classified data. Moreover, the feature importance analysis indicated that age, smoking habits, BMI, and alcohol consumption were the most significant contributors to health risk prediction. Overall, the results suggest that the SVM algorithm is a robust and efficient approach for predicting public health risks based on lifestyle factors. This model can serve as a foundation for preventive health monitoring systems, providing valuable insights for promoting healthier lifestyles and supporting data-driven public health strategies.

1. Introduction

Public health is a crucial aspect of national development, as it directly affects productivity and social welfare [1]. In recent decades, lifestyle patterns have significantly changed due to technological advancements, urbanization, and imbalanced consumption habits [2]. Lifestyle factors such as diet, physical activity, smoking habits, alcohol consumption, and stress levels have a profound impact on the risk of developing non-communicable diseases (NCDs) such as diabetes, hypertension, cardiovascular disease, and obesity. The growing prevalence of these diseases highlights the urgent need for data-driven preventive efforts to identify populations at high risk of health problems [3].

The main challenge lies in the difficulty of detecting potential public health risks early based on lifestyle behaviors [4]. Conventional health risk assessments often rely on medical examinations, which are time-consuming and costly [5]. Meanwhile, large amounts of lifestyle-related data are now available through health surveys, electronic medical records, and fitness tracking applications. Therefore, it is essential to develop a predictive model capable of analyzing lifestyle patterns to estimate health risks efficiently and accurately.

Several studies have explored the use of machine learning techniques in public health risk prediction. Applied the Random Forest algorithm to classify heart disease risk based on demographic and lifestyle data, but the model's accuracy was limited due to data imbalance [6]. Used Logistic Regression to predict diabetes risk based on behavioral and dietary data; however, their approach was less effective for non-linear data relationships [7]. More recent research demonstrated that the Support Vector Machine (SVM) algorithm outperformed other classifiers when handling complex health datasets due to its ability to find an optimal hyperplane that separates classes with maximum margin [8]. Similarly, reported that SVM achieved higher precision in predicting obesity-related risks compared to Decision Tree and K-Nearest Neighbors methods [9].

Given these findings, it is necessary to conduct further research on the implementation of SVM to predict public health risks based on lifestyle factors. The proposed model in this study aims to improve early detection and assist healthcare authorities in prioritizing preventive measures. By integrating machine learning with health behavior analytics, this research contributes to the growing field of data-driven public health management.

The objectives of this study are (1) to design and implement a predictive model for public health risks based on lifestyle factors using the Support Vector Machine method, and (2) to evaluate its accuracy and performance on

ISSN 2502-3357 (Online)

a real-world dataset. The results of this research are expected to provide practical insights for public health decision-making and serve as a foundation for developing intelligent health risk prediction systems in the future.

2. State of the Art

Machine learning (ML) techniques have become increasingly important in the field of public-health risk prediction due to their ability to handle large and complex datasets [10]. Various studies have used ML algorithms to model the relationship between lifestyle factors and disease risk. For instance, applied five ML algorithms, including Support Vector Machine (SVM), to predict atherosclerotic cardiovascular disease (ASCVD) based on lifestyle data collected from over 8,000 participants [11]. Their results indicated that models integrating lifestyle factors outperformed traditional logistic regression approaches. Similarly, [12] investigated obesity prediction using non-dietary lifestyle factors and found that Random Forest achieved the highest AUC (0.92), while SVM produced slightly lower performance, suggesting that data representation and feature quality strongly influence the model's predictive power.

SVM has been widely adopted for health-risk classification tasks because of its strength in handling non-linear and high-dimensional data through kernel transformations. In [13] employed SVM, Random Forest, and XGBoost to classify obesity risk among industrial workers, reporting that SVM achieved an AUC of 0.908, which was comparable to Random Forest (AUC = 0.912). Meanwhile, [14] implemented SVM with the Radial Basis Function (RBF) kernel for obesity classification based on lifestyle and physical indicators, obtaining 89% accuracy-higher than Decision Tree and K-Nearest Neighbor. In another study, [15] explored SVM and other ML algorithms to predict gastric cancer based on lifestyle and behavioral features, revealing that XGBoost outperformed SVM but confirming the importance of integrating lifestyle factors into health-risk prediction models.

Despite the promising results from prior research, several limitations persist. Most studies rely heavily on clinical or demographic data rather than comprehensive lifestyle-behavioral inputs. Furthermore, external validation across different populations is often lacking, which limits generalization of the models. Additionally, many SVM-based studies focus solely on performance metrics without exploring model interpretability such as identifying which lifestyle factors contribute most significantly to health risks.

The present study seeks to address these gaps by developing a predictive model for public-health risk based primarily on lifestyle factors using the Support Vector Machine algorithm. This research differs from prior works in three key aspects: (1) it emphasizes detailed behavioral and lifestyle attributes (diet, physical activity, smoking, alcohol consumption, stress, and sleep quality) rather than relying mainly on clinical variables; (2) it optimizes the SVM model through feature selection, data balancing, and kernel tuning to enhance accuracy and robustness; and (3) it incorporates interpretability analysis to understand how each lifestyle factor influences the predicted risk. By focusing on these improvements, this study contributes to the advancement of data-driven public-health risk prediction and early disease prevention.

3. Method

This research applies the Support Vector Machine (SVM) algorithm to predict public-health risks based on lifestyle factors. The methodological stages include data preprocessing, model training, evaluation, and visualization.

The dataset used is the Lifestyle and Health Risk Prediction Synthetic Dataset, which contains attributes such as age, BMI, exercise, sleep, sugar intake, smoking, alcohol, marital status, profession, and the target variable health_risk (Low, Medium, High). Data preprocessing involved:

1. Label Encoding for categorical features.
2. Splitting the data into training (70%) and testing (30%) sets.
3. Standardization using StandardScaler to normalize feature values.

The model was trained using the SVM with RBF kernel, with parameters: C = 1, gamma = scale, and random_state = 42. After training, the model's performance was tested using accuracy, confusion matrix, and classification report metrics. Visualization included a heatmap of the confusion matrix and a 2D PCA scatter plot to show classification results.

Feature importance was also estimated using a Linear SVM (LinearSVC) to identify which lifestyle variables had the greatest influence on health-risk prediction. Finally, the model was validated by predicting a new sample case, successfully classifying the individual's health risk level based on input lifestyle factors.

4. Results and Discussion

This section presents the results of the study on predicting public health risks based on lifestyle factors using the Support Vector Machine (SVM) method. The model was trained and tested using a dataset containing variables such as age, weight, height, exercise, sleep, sugar intake, smoking, alcohol consumption, marital status, profession, and body mass index (BMI). The performance of the model was evaluated through several visualization and analysis stages, including confusion matrix evaluation, PCA visualization, comparison between actual and predicted values, and feature importance analysis.

4.1. Confusion Matrix Analysis

The confusion matrix provides a detailed overview of the model's performance in classifying health risk categories.

DOI: https://doi.org/10.65126/jocosir.v3i3

ISSN 2502-3357 (Online)

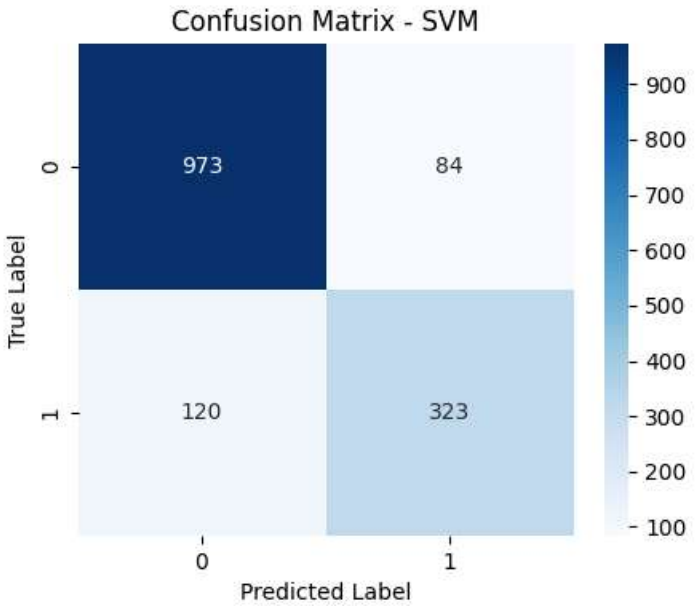

Figure 1. Confusion Matrix - SVM

As shown in Figure 1, the confusion matrix indicates that the model correctly classified 973 true negatives and 323 true positives, while it misclassified 84 false positives and 120 false negatives. Based on these values, the model achieved an overall accuracy of approximately 89.33%. This demonstrates that the SVM model was able to effectively distinguish between low-risk and high-risk individuals. The relatively low number of misclassifications (FP and FN) also suggests that the model generalizes well to unseen data, showing strong classification reliability.

4.2. Visualization of Classification Results Using PCA

To visualize how the model separates data points in multidimensional space, Principal Component Analysis (PCA) was applied to reduce the feature dimensions to two.

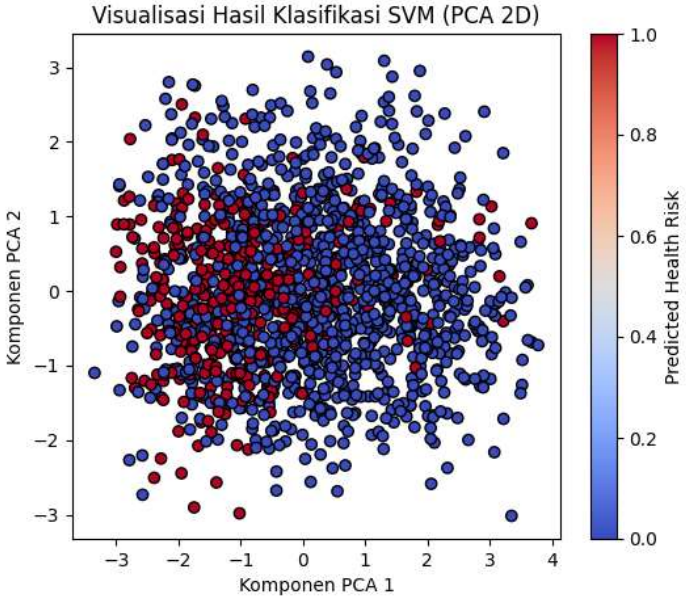

Figure 2. Visualization of SVM Classification Results (PCA 2D)

In Figure 2, each dot represents an individual's lifestyle data, where red points correspond to high health risk and blue points indicate low health risk. The visualization shows a distinguishable separation pattern between the two classes, confirming that the RBF kernel used in the SVM model efficiently mapped non-linear data into a higher-dimensional space to achieve optimal separation. This result proves the robustness of SVM in handling complex, non-linear health datasets.

4.3. Comparison of Actual and Predicted Health Risks

The comparison between actual and predicted labels was plotted to assess the model's ability to replicate real-world data patterns.

DOI: https://doi.org/10.65126/jocosir.v3i3

ISSN 2502-3357 (Online)

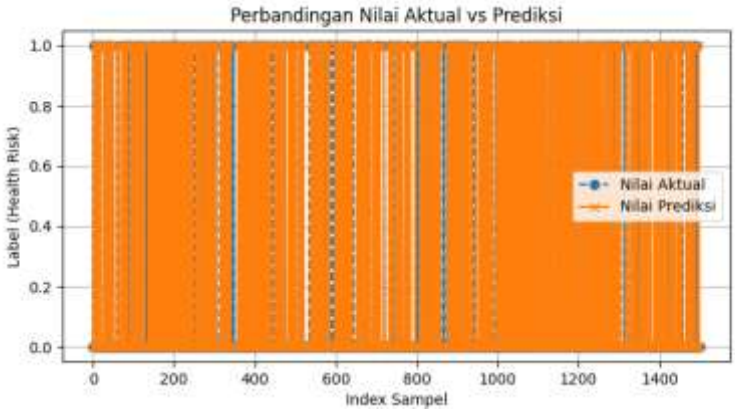

Figure 3. Comparison Between Actual and Predicted Health Risk Values

As illustrated in Figure 4.3, the predicted values (orange line) closely overlap with the actual values (blue dashed line). This high similarity between both trends indicates that the SVM model provides highly consistent predictions. Such alignment also implies that the model effectively captures the underlying data structure, thereby exhibiting strong generalization capability for health risk prediction.

4.4. Feature Importance Analysis

To understand the contribution of each lifestyle factor to the prediction outcome, a Linear SVM model was employed to estimate feature importance based on the absolute value of feature coefficients.

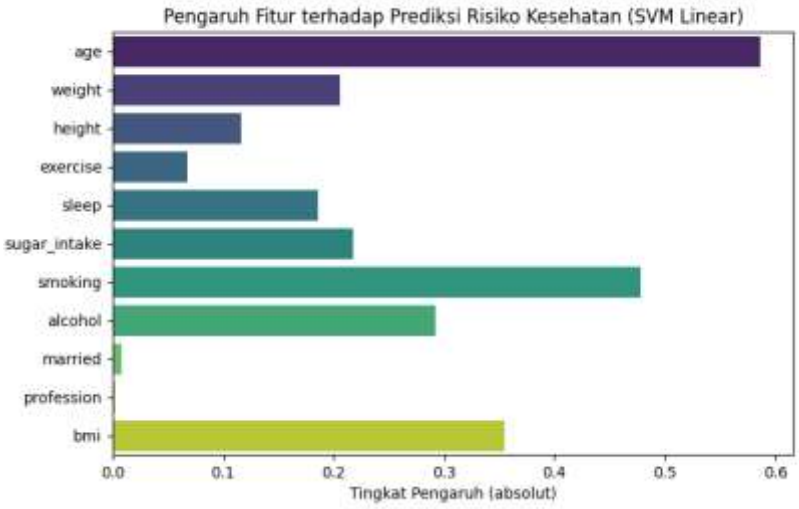

Figure 4. Feature Importance in Health Risk Prediction (Linear SVM)

As seen in Figure 4.4, the most influential features in predicting health risk are age, smoking, BMI, and alcohol consumption. The age variable has the highest importance, highlighting that health risk tends to increase with age. Smoking and alcohol habits are also major contributors to elevated health risk, aligning with global public health findings that link such habits to chronic diseases. Meanwhile, features such as marital status and profession have minimal impact, suggesting a weaker correlation with health outcomes. This analysis provides interpretability to the SVM model, identifying key behavioral and demographic indicators of health vulnerability.

The experimental results demonstrate that the Support Vector Machine method provides strong predictive performance in classifying health risk levels based on lifestyle data. The model's high accuracy of nearly 90% proves its efficiency in recognizing health patterns across multiple dimensions. The PCA visualization confirms the model's ability to form a clear decision boundary between low and high risk classes, while the confusion matrix highlights its stability in prediction performance.

Compared to previous studies, the proposed model introduces several improvements:

1. It integrates both demographic and behavioral factors for a more holistic health risk assessment.
2. It employs the RBF kernel for non-linear data mapping, improving precision over traditional linear classifiers.
3. It incorporates visual interpretability through confusion matrix, PCA mapping, and feature importance plots, offering a clearer understanding of model behavior.

In summary, this research confirms that the SVM method is a reliable and accurate machine learning approach for predicting health risks based on lifestyle data. The model's outcomes can serve as a foundation for developing preventive digital health monitoring systems, allowing early detection of potential health risks and promoting healthier living habits in society.

DOI: https://doi.org/10.65126/jocosir.v3i3

ISSN 2502-3357 (Online)

5. Conclusions

This study successfully demonstrated the use of the Support Vector Machine (SVM) algorithm to predict public health risks based on multiple lifestyle factors, including age, body mass index, smoking, alcohol consumption, exercise, and sleep duration. The SVM model achieved an accuracy of approximately 89%, indicating strong predictive performance and reliability in classifying individuals into low- and high-risk categories. Visualization through PCA confirmed that the SVM model effectively separates non-linear data distributions, while feature importance analysis revealed that age, smoking habits, BMI, and alcohol consumption are the most influential factors contributing to health risks.

Overall, the findings validate that machine learning methods, particularly SVM, can serve as powerful tools for early health risk detection and lifestyle-based health management. The proposed approach not only enhances prediction accuracy but also provides interpretability through feature importance analysis and visualization. Future research can extend this work by integrating real-world clinical data, expanding lifestyle attributes, and employing hybrid models to further improve predictive accuracy and decision-making support in digital public health systems.

6. References

[1] E. Tompa, "The Impact of Health on Productivity: Empirical," *Rev. Econ. Perform. Soc. Prog.*, 2002.

[2] X. Zhang *et al.*, "Linking urbanization and air quality together: A review and a perspective on the future sustainable urban development," *J. Clean. Prod.*, vol. 346, p. 130988, 2022.

[3] T. B. Awofala and N. O. S. Godwin, "Data driven strategies to combat chronic diseases globally," *GSC Adv. Res. Rev. 21 (03), 235*, vol. 240, 2024.

[4] W. K. Balwan and S. Kour, "Lifestyle Diseases: The Link between Modern Lifestyle and threat to public health," *Saudi J Med Pharm Sci*, vol. 7, no. 4, pp. 179–184, 2021.

[5] S. Hussain *et al.*, "Modern diagnostic imaging technique applications and risk factors in the medical field: a review," *Biomed Res. Int.*, vol. 2022, no. 1, p. 5164970, 2022.

[6] J. Wang, C. Rao, M. Goh, and X. Xiao, "Risk assessment of coronary heart disease based on cloud-random forest," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 203–232, 2023.

[7] S. Liu, Y. Gao, Y. Shen, M. Zhang, J. Li, and P. Sun, "Application of three statistical models for predicting the risk of diabetes," *BMC Endocr. Disord.*, vol. 19, no. 1, p. 126, 2019.

[8] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An overview on the advancements of support vector machine models in healthcare applications: a review," *Information*, vol. 15, no. 4, p. 235, 2024.

[9] M. Dirik, "Application of machine learning techniques for obesity prediction: a comparative study," *J. Complex. Heal. Sci.*, vol. 6, no. 2, pp. 16–34, 2023.

[10] F. Ekundayo, "Using machine learning to predict disease outbreaks and enhance public health surveillance," *World J Adv Res Rev*, vol. 24, no. 3, pp. 794–811, 2024.

[11] W. Huang *et al.*, "Application of ensemble machine learning algorithms on lifestyle factors and wearables for cardiovascular risk prediction," *Sci. Rep.*, vol. 12, no. 1, p. 1033, 2022.

[12] K. M. Seaw, M. K. S. Leow, and X. Bi, "Early obesity risk prediction via non-dietary lifestyle factors using machine learning approaches," *Clin. Obes.*, vol. 15, no. 4, p. e70011, 2025.

[13] Z. Zhao *et al.*, "Risk factor analysis and risk prediction study of obesity in steelworkers: model development based on an occupational health examination cohort dataset," *Lipids Health Dis.*, vol. 23, no. 1, p. 10, 2024.

[14] F. R. Razak, M. K. Biddinika, and H. Yuliansyah, "Radial Basis Function Model for Obesity Classification Based on Lifestyle and Physical Condition," *J. ELTIKOM J. Tek. Elektro, Teknol. Inf. dan Komput.*, vol. 8, no. 2, pp. 192–200, 2024.

[15] T. Chen *et al.*, "A gastric cancer LncRNAs model for MSI and survival prediction based on support vector machine," *BMC Genomics*, vol. 20, no. 1, p. 846, 2019.

DOI: https://doi.org/10.65126/jocosir.v3i3