Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 4 No. 2 (2025) pp: 4855-4860
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Analysis of Tobacco Expenditure Patterns and Health in Five Provinces of Indonesia: Descriptive and Segmentative Approaches

Syafiq Muhammad Fauzi [1], Zurnan Alfian[2]
[1]Fakultas Ilmu Komputer, Prodi Teknik Informatika, Universitas Pamulang, Kota Tangerang Selatan, Indonesia.
Email: [1]fieqmuhammad8@gmail.com

Abstract

This study aims to identify and analyze community consumption patterns of tobacco products and health expenditures based on 1999 data released by the Indonesian Central Statistics Agency (BPS). The study focuses on five representative provinces in Indonesia: Aceh, Riau, Bengkulu, Bali, and North Sulawesi. Using a descriptive approach and the K-Means Clustering data mining algorithm, this research successfully maps two main consumption patterns that show a negative correlation between tobacco expenditures and the allocation of funds for health needs. Practically, these results provide a cross-regional segmentative overview that can serve as a basis for formulating more responsive public policies, particularly in the areas of health and tobacco consumption control. This study also highlights the importance of integrating historical data with modern analytical approaches to support evidence-based decision-making processes.

Keywords: Tobacco Consumption; Health Expenditure; Data Mining; K-Means Clustering; Public Policy; Indonesia.

Introduction

Community consumption behavior is a real representation of the social, cultural, and economic conditions that develop in a region. In the context of national development, understanding the structure and priorities of consumption is crucial, especially in formulating public policies that are responsive to community needs. One important aspect of consumption behavior is spending on goods that have a direct impact on health, such as tobacco products, and the allocation of funds for health needs themselves.

In Indonesia, the phenomenon of tobacco consumption is still a serious issue. Reports from the Central Statistics Agency and various public health studies show that the level of cigarette consumption in various provinces tends to be high even though various regulatory and public education efforts have been made. Ironically, the high expenditure on tobacco products is often not accompanied by a balanced allocation of expenditure for health needs such as routine medical check-ups and health insurance. This imbalance can worsen the burden on public health in the long term, especially in areas with low income levels.

In 1999, the Central Statistics Agency released per capita consumption data that included the allocation of community expenditure for various basic needs including tobacco and health. The five provinces analyzed in this study, namely Aceh, Riau, Bengkulu, Bali, and North Sulawesi, showed different consumption dynamics that could reflect the level of health

awareness, the influence of local culture, and access to health services. For example, provinces with strong smoking culture values tend to have a larger expenditure allocation for tobacco while provinces with better health infrastructure tend to allocate more funds for medical needs.

In addition, the level of education and public health literacy are also important factors that influence the priority of public consumption towards tobacco and health services. People with higher levels of education tend to have better awareness in maintaining health and reducing consumption of tobacco products. However, in some areas with limited access to education and health information, people still consider smoking as part of social culture so that it is difficult to reduce consumption levels.

To understand more deeply the characteristics of consumption behavior, a data mining approach can be utilized, especially unsupervised learning techniques such as K Means Clustering. This method allows data grouping based on similarities in consumption patterns so that regional segments with certain behavioral characteristics can be identified. In other words, this algorithm provides useful insights to explore the potential relationship between public consumption preferences for tobacco and health awareness levels.

This study seeks to fill the existing knowledge gap by integrating an algorithmic approach based on data mining with a theoretical basis for consumption behavior. The results of this grouping are expected to contribute to policy makers both at the central and regional levels in designing more targeted intervention strategies such as anti-smoking campaigns, health education or increasing access to medical services in areas with low health spending. By combining descriptive and segmentative analysis methods, this study not only serves as a retrospective analysis of 1999 data but also as a starting point for the development of predictive models in analyzing future consumption behavior. The use of historical data with modern methods allows the creation of a more accurate tobacco consumption and health segmentation map, thereby assisting the government and health institutions in formulating evidence-based policies. In the future, the results of this study can also be used as a basis for identifying provinces that require more intensive intervention to control tobacco consumption and improve health services in order to create a healthier and more productive society.

Research Method

This study is an exploratory quantitative study using a data mining approach with the K-Means Clustering algorithm to explore tobacco consumption patterns and public health conditions in five Indonesian provinces (Aceh, Riau, Bengkulu, Bali, and North Sulawesi) based on 1999 data from the Indonesian Central Statistics Agency (BPS). The study uses two main variables: the percentage of expenditure on tobacco (X1) and the percentage of expenditure on health (X2), chosen for their relevance in depicting consumption behavior and public health issues.

This approach is utilized to identify cluster patterns within data that are not explicitly labeled, allowing the results to map regions with high tobacco expenditures and low health allocations. The findings are expected to serve as a basis for formulating policies on tobacco consumption control and increasing public health awareness, while providing insights into the relationship between consumption preferences and the socioeconomic conditions of each province.

Analysis Stages:

- **Pre-processing:** Data from the official statistics agency were prepared using min-max scaling normalization to ensure each variable is within a 0–1 scale, allowing tobacco and health expenditure variables to have equal weighting in the clustering process. This step is essential to prevent variables with larger scales from dominating and influencing the cluster formation results.

DOI: https://doi.org/10.31004/riggs.v4i2.1341

- **Cluster Determination:** The number of clusters (k) was set to two based on initial data exploration and scatter plot visualization, which indicated a natural separation among the provinces. Choosing k=2 also considers the policy analysis context, making it easier to translate findings into actionable public health intervention recommendations.
- **Clustering:** The K-Means algorithm was applied using Euclidean distance to measure the similarity between data points until the centroid positions stabilized. The algorithm iteratively reassigns cluster members based on proximity to new centroids until no significant reassignment occurs, resulting in a provincial division based on tobacco and health expenditure patterns.
- **Visualization and Evaluation:** Clustering results were visualized using scatter plots with different colors for each cluster, facilitating easy identification of patterns and outliers. The evaluation was conducted by linking the findings with each region's socioeconomic conditions; for example, areas with high tobacco spending but low health allocation may require aggressive policies, while areas with high health expenditures may focus on strengthening preventive health services.

Tools Used:

- Excel/Google Sheets were used for initial data input, double-data validation, and manual outlier checking to ensure data cleanliness before further processing.
- Python (Jupyter Notebook) with pandas, sklearn, matplotlib, seaborn, and numpy libraries was utilized for data processing (scaling and cleaning), K-Means Clustering, and result visualization through scatter plots and distribution diagrams.
- By integrating Python, the analysis can be automatically repeated when data updates occur, supporting reproducible research practices and systematic documentation of the analysis pipeline.

The data were sourced from the official statistics agency (BPS), ensuring validity. Although historical, the data remain relevant as a foundation for examining cross-regional patterns in tobacco consumption and health expenditure in Indonesia. This is crucial as an initial reference to support data-driven public health planning and policy-making, especially in identifying priority regions for anti-tobacco campaigns, improving basic health infrastructure, and strengthening targeted preventive health services.

Results and Discussion

Based on the data processing results, two main clusters were identified with the following characteristics. These findings are an essential part of the analysis as they help us understand how community consumption patterns, particularly regarding tobacco and health expenditures, can form different regional segmentations. This segmentation not only provides a visual representation of each region's condition but also serves as a basis for further evaluation of previously implemented policies.

Jenis Pengeluaran	Daerah	Aceh (%)	Riau (%)	Bengkulu (%)	Bali (%)	Sulut (%)
Pengeluaran untuk makanan						
Tembakau, Rokok	Desa	6.18	6.57	6.77	2.92	5.24
	Kota	5.29	4.62	4.63	2.18	3.89
	Desa+Kota	5.92	5.66	6.08	2.60	4.68
Pengeluaran non makanaan						
Kebutuhan Kesehatan	Desa	1.60	1.59	1.10	2.19	1.74
	Kota	1.99	1.98	2.02	2.23	2.16

DOI: https://doi.org/10.31004/riggs.v4i2.1341

	Desa+Kota	1.71	1.75	1.37	2.21	1.91

Cluster 0 – High Tobacco Consumption, Low Health Expenditure

- **Provinces:** Aceh, Riau, Bengkulu
- **Average tobacco expenditure:** > 6.5%
- **Health expenditure:** < 2%
- **Indication:** Strong smoking culture, low awareness of health services, and potentially limited medical access.

Cluster 1 – Higher Health Expenditure, Lower Tobacco Consumption

- **Provinces:** Bali, North Sulawesi
- **Average health expenditure:** > 2%
- **Tobacco expenditure:** < 6%
- **Indication:** Better health awareness, more evenly distributed medical infrastructure, and possibly better-controlled smoking culture.

Visualization of the clustering results shows a clear distinction between these two groups. This is evident from the consistent data distribution patterns, where provinces within the same cluster have a higher percentage of tobacco expenditure but lower health allocation compared to those in the other group. This phenomenon provides a clear picture of how community consumption behavior can affect their aggregate health quality.

In terms of policy, these findings serve as an important indication that health education and preventive approaches should be tailored to each region's characteristics. Provinces within Cluster 0 require more aggressive intervention strategies, including anti-tobacco campaigns and improvements in basic health infrastructure. Meanwhile, regions classified under Cluster 1 can focus on strengthening preventive health services and enhancing sustainable community access to health programs.

Furthermore, this segmentative approach opens opportunities for developing AI- and data mining-based predictive models for regional planning and social budget allocation. Through such modeling, local governments can map potential health risks based on community consumption patterns, enabling interventions to be conducted more efficiently and in a targeted manner, aligned with historical data and spending trends. These predictive models can utilize supervised learning algorithms such as regression or random forest to predict regions with high health risk potential due to tobacco consumption while identifying demographic and economic factors contributing to these elevated risks.

Additionally, the integration of AI and data mining in public policy enables governments to allocate health budgets more precisely by considering regional clusters and predicted future disease burdens. This approach can be used to plan health education campaigns, strengthen preventive services, and optimize the distribution of healthcare facilities according to each region's needs. Thus, policies implemented become not only reactive but also proactive, minimizing budget waste by ensuring each intervention is data-driven and relevant to local community conditions.

Moreover, the use of clustering algorithms on tobacco and health expenditure data can serve as an early indicator for mapping potential regional vulnerabilities to chronic health issues. Fiscal policies and health budget distribution can be directed more

DOI: https://doi.org/10.31004/riggs.v4i2.1341

equitably based on regional conditions, thereby increasing the effectiveness of public fund utilization and delivering tangible benefits to the community.

This data analysis integration also assists academics, researchers, and government agencies in identifying priority areas for public health interventions. Data mining plays an active role beyond passive analysis, enabling the monitoring of changes in consumption and health patterns, allowing policy evaluations to be conducted periodically to ensure interventions remain relevant to societal dynamics. Therefore, this study makes a tangible contribution to evidence-based policy practices. This approach is crucial not only in the context of tobacco consumption control but also applicable to other public health issues requiring a segmentative, data-driven approach, enabling policies to be more targeted and effective in improving community welfare across various regions in Indonesia.

Conclusion

This study utilized the K-Means Clustering algorithm to analyze per capita expenditure patterns on tobacco and health across five provinces using 1999 data from the Indonesian Central Statistics Agency (BPS). The analysis revealed two main clusters: (1) provinces with high tobacco expenditure and low health expenditure, namely Aceh, Riau, and Bengkulu; and (2) provinces with the opposite pattern, namely Bali and North Sulawesi. This segmentation indicates that disparities in consumption priorities persist in several regions, likely influenced by cultural factors, smoking habits, and the level of health awareness. Regions with high tobacco expenditure and low health allocation can serve as early indicators for policy intervention, particularly in health education and tobacco consumption control. The findings also demonstrate that clustering methods can provide a straightforward yet informative overview of consumption patterns across regions. Although the data used is historical, this approach remains relevant as a foundation for formulating data-driven policies in regions with similar characteristics. Future research is recommended to utilize more recent data and consider additional variables to enrich segmentation and behavioral consumption analysis. This study confirms that data analysis approaches using the K-Means Clustering algorithm can be effectively employed to uncover community consumption patterns regarding tobacco and health across various regions in Indonesia. By using historical data from BPS in 1999, this study not only offers objective regional classifications but also provides insights into how consumption behavior is influenced by cultural factors, health awareness, and access to public services. Moving forward, the integration of data mining into social policy planning should continue to be encouraged, not only as an analytical tool but also as a driver for evidence-based policy transformation. It is hoped that central and local governments can utilize findings such as these to design more targeted intervention programs, particularly in regions with high health risks but low awareness. Ultimately, the results of this research are expected to serve as an initial foundation for encouraging broader follow-up studies, in terms of the number of provinces analyzed, the diversity of variables, and the integration of advanced analytical technologies to address increasingly complex public health challenges.

References

BPS. (1999). *Per Capita Household Consumption in Indonesia.* Jakarta: BPS.

Han, J., Kamber, M., & Pei, J. (2013). *Data Mining: Concepts and Techniques (3rd ed.).* Burlington: Morgan Kaufmann.

Ministry of Health, R. o. (2010). *National Action Plan for the Control of Tobacco Consumption Impact.* Jakarta: KemenKes RI.

Ministry of Health, R. o. (2013). *Basic Health Research (Riskesdas) 2013.* Jakarta: National Institute of Health Research and Development.

Nurcahyo, H., & Fadjar, H. (2016). Data Mining in the Health Sector: Application of Clustering to Analyze Health Behavior. *Journal of Information Technology and Computer Science, 3(1)*, 23-30.

Organization, W. H. (2011). *WHO Report on the Global Tobacco Epidemic: Warning about the Dangers of Tobacco.* Geneva: World Health Organization.

DOI: https://doi.org/10.31004/riggs.v4i2.1341

Rokhmah, D. (2012). Smoking Consumption Behavior and Its Impact on Public Health. *Journal of Public Health*, 115-123.

Statistics, C. B. (2010). *Indonesia Health Statistics.* Jakarta: BPS.

Suhartono, R. (2013). *Data Mining: Theory and Applications.* Bandung: Informatika.

Sulaiman, S. (2009). *Understanding the Use of Economic Science in Health Management.* Yogyakarta: Gadjah Mada University Press.

DOI: https://doi.org/10.31004/riggs.v4i2.1341