Utilization of K-means Clustering for Classifying Diabetes Risk Populations According to Health Behaviors and 3Es-2Ss Health Literacy

Supaporn Yodmunee[1], Wongpanya S. Nuankaew[2], Thapanapong Sararat[3], Pratya Nuankaew[4,*]

[1]*Ngao District Health Office, Lampang, 52110, Thailand*

[2]*Department of Computer Science, School of Information and Communication Technology, University of Phayao, Phayao, 56000 Thailand*

[3]*Department of Computer Graphics and Multimedia, School of Information and Communication Technology, University of Phayao, Phayao, 56000 Thailand*

[4]*Department of Digital Business, School of Information and Communication Technology, University of Phayao, Phayao, 56000 Thailand*

(Received: October 4, 2025; Revised: December 1, 2025; Accepted: March 1, 2026; Available online: April 4, 2026)

Abstract

This study focused on classifying populations at risk for diabetes using K-means clustering integrated with the 3Es–2Ss health literacy framework: eating, exercise, emotion, smoking cessation, and alcohol cessation. Biological, behavioral, and health literacy data were analyzed. The dataset was collected from 126 participants identified as at-risk individuals in Ngao District, Lampang Province, Thailand. This relatively small, community-based sample provides valuable insights into local health behaviors but limits the generalizability and statistical power of the findings to broader populations. The K-means clustering analysis, guided by the Elbow method, identified k = 4 as the optimal number of clusters, yielding four distinct groups with different socio-demographic and health characteristics. These clusters revealed variations in health profiles, economic status, and behavioral literacy within the Thai population. Despite the small sample size and limited generalizability, missing data and inconsistencies were systematically addressed through data cleaning and normalization to maintain analytical reliability. The results suggest that K-means clustering can serve as an effective decision-support tool for public health planning, particularly for Non-Communicable Disease (NCD) prevention and diabetes management at the local level.

Keywords: 3Es–2Ss Health Literacy, Diabetes Risk Classification, Diabetes Risk Populations, Health Behaviors, NCD Prevention

1. Introduction

Diabetes mellitus is becoming a growing public health problem worldwide, with predictions indicating there could be up to 783 million cases by 2045 [1]. This burden not only strains healthcare systems but also worsens the quality of life for those affected [1]. Epidemiological data in Thailand show a consistent rise in the prevalence of type 2 diabetes, especially among working-age adults and the elderly, primarily due to unhealthy habits and low health literacy [2], [3]. For effectiveness, prevention efforts must target both biological risk factors, such as weight, waist circumference, and fasting blood sugar, and health behaviors, including nutrition, exercise, emotional regulation, and the use of alcohol and tobacco simultaneously.

Thailand developed the "3Es–2Ss framework," referring to Eating, Exercise, Emotion, Stop Smoking, and Stop Alcohol, as a central model for promoting health literacy and preventing NCDs [4]. International data confirms a strong link between health literacy and both diabetes self-management and clinical outcomes [5]. This approach enables culturally relevant assessments that ensure consistency across diverse global populations.

The data files collected from questionnaires in this project accurately represent risk factors related to diabetes, including biological data such as gender, age, weight, height, waist circumference, and fasting blood sugar levels; health behaviors like dietary habits, exercise, emotional regulation, alcohol use, and smoking; and social factors such as

***Corresponding author: Pratya Nuankaew (pratya.nu@up.ac.th)**
DOI: https://doi.org/10.47738/jads.v7i2.1042

participation in health-related activities, community involvement, satisfaction with healthcare providers, and access to information. This supports the idea that preventing chronic diseases requires addressing biological, behavioral, and social aspects together [6]. In public health research, "traditional analysis" usually involves descriptive statistics, hypothesis testing, and logistic regression to explore risk factors for diabetes. These methods struggle with multidimensional, nonlinear data like biological traits, health behaviors, and health literacy, making it hard to identify "risk subgroups."

Traditional public health studies typically employ descriptive statistics, hypothesis testing, and logistic regression to identify diabetes risk factors. While appropriate for linear, low-dimensional data, these methods fall short when handling the multidimensional, nonlinear nature of health data such as those derived from the 3Es–2Ss framework, which blends behavioral, literacy, and biological components. K-means clustering fills this gap by clustering populations based on data similarity without pre-specifying target variables, and can reveal hidden risk structures that may not be visible to traditional statistical methods.

That is, traditional methods help answer the question “What factors influence diabetes risk?”, while K-means clustering helps answer deeper questions such as “Which subpopulations have unique risk patterns?”, which is useful for more targeted public health policy and interventions.

K-means clustering automatically classifies populations based on data similarity, revealing data structures often missed by traditional methods. Therefore, articles should clarify that "traditional analysis" refers to basic statistics and logistic regression to highlight the difference from K-means in detecting hidden risk groups. Unsupervised machine learning, specifically K-means clustering, has been used for comprehensive risk stratification [7]. Studies in Europe and Asia show that K-means clustering can effectively classify diabetic patients and their risk groups, uncovering clinically significant diabetic sub-clusters with different metabolic features and treatment needs [8].

In summary, traditional methods answer “Which factors influence diabetes risk?”, while K-means clustering addresses “Which subpopulations share unique risk profiles?”, a distinction crucial for designing targeted interventions—particularly within the 3Es–2Ss framework that integrates both physical and mental health behaviors.

This research applied the K-means clustering technique to questionnaire data from the Thai population, using the 3Es–2Ss framework as the core of the analysis. This framework reflects key health behaviors of Thais in a real-world context: eating, exercise, emotion, quitting smoking, and quitting alcohol. Each component plays a specific role in determining diabetes risk behaviors. Eating foods that are too sweet, fatty, or salty is a significant cause of insulin resistance. Insufficient exercise reduces energy expenditure, while emotion management influences eating and sleeping behaviors. Smoking and drinking not only increase the risk of cardiovascular disease but also affect blood sugar control. Therefore, the 3Es–2Ss framework provides a comprehensive picture of health behaviors associated with diabetes risk across the biological and social dimensions of Thais [8].

To ensure accurate results, this research used the elbow method to normalize the data and determine the appropriate number of groups. It repeatedly initialized the K-means model to verify its stability [9]. Each group was then externally validated using metabolic and health literacy indicators to characterize the groups. For example, groups with low health literacy and inappropriate health behaviors may require community-level promotion. In contrast, groups with good behaviors but high metabolic risk may require closer care from primary care facilities [6], [8].

Overall, using field data in conjunction with the 3Es–2Ss conceptual framework and K-means clustering techniques enables more accurate classification and understanding of "diabetes risk groups." The clustering results are expected to reflect risk trends in the Thai population at the local level, supporting policy planning focused on diabetes prevention and control. They can also be developed as practical decision-support tools for primary health care management.

The primary objective of this study is to develop and implement K-means clustering for classifying diabetes risk populations by integrating biological, behavioral, and health literacy data within the 3Es–2Ss framework. The focus is to identify latent risk subgroups—hidden population structures that traditional statistical analyses cannot detect—and to utilize these insights to guide community-level and policy-level diabetes prevention strategies. The secondary objectives are clarified into three distinct roles.

(1) Data Integration (Biological–Behavioral–Literacy): To collect and interlink biological variables (e.g., age, BMI, waist circumference, fasting blood sugar) with health behavior variables (eating, exercise, emotion regulation, smoking cessation, and alcohol cessation) and health literacy measures within the 3Es–2Ss framework, forming a comprehensive dataset that reflects Thai health behavior patterns. (2) Analytical Application: To apply K-means clustering for categorizing populations based on diabetes risk, and to validate the model's accuracy, stability, and performance using statistical methods such as the Elbow Method, Silhouette Score, and Davies–Bouldin Index. (3) Interpretation and Policy Insight: To interpret and describe each identified cluster by linking metabolic, behavioral, and literacy characteristics, providing actionable insights for targeted public health strategies and diabetes prevention policies tailored to specific subgroups.

2. Literature Reviews

2.1. The 3Es–2Ss Framework in Thai Health Research

Over the past five years, the 3Es–2Ss framework—comprising Eating, Exercise, Emotion, Stop Smoking, and Stop Alcohol—has become a central tool in Thai health research, influencing both policy and practical applications. For instance, Thongsong and Neranon [10] developed an obesity prevention literacy scale based on this framework, Pitchalard et al. [11] applied it to train village health volunteers in NCD management, and Turnbull et al. [12] integrated it with artificial intelligence to predict undiagnosed hypertension among rural Thai elders. These studies demonstrate how the 3Es–2Ss framework has evolved beyond behavior promotion to serve as an analytical and technological foundation for modern Thai health systems.

Nevertheless, despite its wide adoption, several research limitations remain unaddressed. First, most studies using the 3Es–2Ss framework focus primarily on individual-level behavior assessments—such as eating and exercise—without adequately integrating biological or metabolic indicators that reflect measurable disease risk. Second, health literacy measurement under this framework often relies on self-reported questionnaires, which may lack reliability and comparability across populations. Third, there is a clear gap in the application of 3Es–2Ss data to advanced analytical methods, such as predictive modeling or machine learning approaches, that could better capture complex nonlinear relationships between behaviors and biological risk factors.

Therefore, integrating the 3Es–2Ss framework with modern computational techniques like K-means clustering can help identify "latent risk subgroups" that traditional linear or regression-based analyses overlook. This integration bridges methodological gaps and enhances the framework's analytical depth, enabling it to inform more precise, data-driven health interventions and policy planning tailored to specific population behaviors and risks.

In summary, while the 3Es–2Ss framework has significantly advanced health literacy and behavioral change initiatives in Thailand, its current applications remain limited by their descriptive focus and lack of integration with data-driven analysis. The present study addresses this gap by merging the 3Es–2Ss framework with K-means clustering to produce more systematic, accurate classifications of diabetes risk groups—thereby contributing both methodological innovation and practical relevance to Thai public health research.

2.2. Burden of Behavioral Risk Factors and Health Literacy

The rising prevalence of Type 2 Diabetes (T2D) underscores the intricate interplay between biological and behavioral risk factors, particularly in Thailand and the broader Southeast Asian region. Epidemiological studies have shown that diets high in fat and salt, coupled with low vegetable consumption, are strongly associated with an increased risk of T2D [13]. While such findings emphasize behavioral patterns, a deeper understanding requires consideration of the underlying biological mechanisms that explain how these behaviors contribute to disease onset.

High intake of sugar and fat stimulates excessive insulin secretion as the body strives to maintain normal blood glucose levels. Over time, this persistent demand leads to insulin resistance—the hallmark of type 2 diabetes. Physical inactivity compounds this effect by reducing the muscles' ability to utilize glucose efficiently, resulting in chronic hyperglycemia. Additionally, smoking and alcohol consumption disrupt pancreatic function and metabolic regulation. Nicotine and ethanol contribute to chronic inflammation and oxidative stress, both of which impair insulin signaling and promote metabolic dysfunction.

Psychological factors also play a biological role. Chronic stress elevates cortisol levels, triggering hepatic gluconeogenesis (glucose production in the liver), which sustains higher blood sugar levels. Consequently, emotional regulation and mental health management are physiologically linked to diabetes prevention—on par with diet and exercise interventions.

Multiple studies have also confirmed that health literacy significantly mitigates behavioral risk factors, as diabetes-specific literacy programs can reduce HbA1c levels and improve self-management practices [14]. However, disparities related to age, education, income, and digital access continue to limit equitable health information dissemination [15].

Therefore, linking behavioral risk factors to their biological underpinnings strengthens the argument that unhealthy habits are not merely statistical correlates but mechanistic contributors to disease progression. Integrating behavioral, literacy, and physiological perspectives is essential for developing sustainable, evidence-based diabetes prevention strategies.

2.3. Unsupervised Learning to Reveal Significant T2D Subgroups

Unsupervised learning techniques, especially K-means clustering, have gained global recognition for their ability to uncover biological and clinical heterogeneity in T2D. A Nature study demonstrated that global Genome-Wide Association Studies (GWAS) could classify mechanistic clusters linked to metabolic dysregulation [13]. Similarly, large-scale cohort studies such as the UK Biobank (n≈420,000) employed K-means and related algorithms to identify reproducible T2D subtypes characterized by differences in mortality, comorbidities, and treatment responses [16], [17]. These studies confirm K-means clustering as an effective tool for identifying metabolic and behavioral subgroups within complex datasets.

However, directly referencing these large-scale global studies risks overstating their applicability to small-scale rural contexts such as Ngao District, Lampang Province, Thailand, where population structures, healthcare access, and socioeconomic conditions differ substantially. Global epidemiological models are typically based on dense, longitudinal datasets and well-resourced healthcare systems—conditions that rarely exist in rural Thailand. This creates an applicability gap, as global findings cannot be directly transferred to community-level health systems without contextual adjustment.

The present study therefore adopts K-means clustering through a localized adaptation approach, using community-level survey data that represent Thailand's biological, behavioral, and literacy characteristics rather than large-scale genomic or hospital-based datasets. The goal is to identify context-specific risk clusters that may be invisible to conventional analyses and to generate findings that can inform local public health strategies, particularly within primary care units and subdistrict health-promoting hospitals.

This approach reframes K-means clustering as a bridge between global analytical frameworks and local healthcare realities, emphasizing practical insight over generalization. Rather than replicating the scale of global studies, this research applies the same analytical logic to smaller, behaviorally rich datasets—transforming machine learning from a purely technical tool into a practical framework for rural health understanding.

In summary, while global studies provide methodological rigor and theoretical depth, their direct application to Thailand's rural health system is limited. By adapting these analytical principles to small-scale, resource-limited environments, this study illustrates how global scientific methods can be translated into actionable, community-level health intelligence.

2.4. Methodological Guardrails for K-means and Validation

Robust segmentation depends on transparent validation. A recent large-scale benchmark compared 68 Cluster-Validity Indices (CVIs) across both real and simulated datasets and provided practical guidance on choosing the best number of clusters and evaluating K-means results in high-dimensional settings [18]. Complementary reviews compile advances and limitations for internal versus external CVIs, highlighting their sensitivity to noise and feature scaling—key considerations when integrating anthropometrics, glycemia, behaviors, and HL metrics [19]. Consequently, studies applying K-means on 3Es–2Ss–anchored survey data should include normalization, multiple random starts, CVI-guided model selection, and external profiling of clusters using metabolic markers and literacy measures to ensure

stability, interpretability, and usefulness [16]. This evidence supports using unsupervised clustering as a decision-support layer in Thai primary care: guiding community-level behavior/HL interventions toward "low-HL/high-risk-behavior" clusters while increasing metabolic monitoring for "good-behavior but high-metabolic-risk" clusters [13].

The reviewed literature emphasizes that type 2 diabetes results from a complex interaction of biological, behavioral, and health literacy factors, all of which influence the success of prevention and management strategies. The 3Es–2Ss framework provides a culturally relevant way to assess key behavioral risks, while diabetes-specific health literacy programs have been shown to improve clinical outcomes and patient self-care. Simultaneously, advances in unsupervised machine learning, especially K-means clustering, have identified consistent subgroups of T2D that differ in prognosis, comorbidities, and care needs, highlighting the potential of data-driven segmentation for precision prevention. However, methodological rigor remains crucial, including normalization, CVI-guided cluster selection, and external validation, to ensure stability and clarity of results. Overall, these findings suggest that combining biological, behavioral, and health literacy factors with robust clustering methods can develop a valuable decision-support tool, enabling more targeted and sustainable diabetes prevention efforts in Thai primary healthcare and community settings.

3. Materials and Methods

3.1. Research Framework and Design

This research framework shows how to classify diabetes risk groups. It starts with biological data like age, gender, weight, height, waist size, and fasting blood sugar, along with health behaviors such as diet, physical activity, emotional management, smoking, and alcohol intake. Health literacy is assessed through the 3Es–2Ss framework, covering Eating, Exercise, Emotion, Stop Smoking, and Stop Alcohol. These are analyzed with K-means clustering to group populations into meaningful sub-risk categories important for public health. The results include diabetes risk profiles that distinguish low health literacy and unhealthy lifestyles from healthier habits with metabolic issues. These insights support targeted prevention strategies at primary healthcare and community levels. The goal is to develop a data-driven tool to improve diabetes management and prevention in Thailand.

The conceptual framework of this study illustrates the integration of biological, behavioral, and health literacy data within the 3Es–2Ss framework into an analytical process using K-means clustering to classify populations by diabetes risk level (figure 1). The analysis begins with biological factors such as age, gender, weight, height, waist circumference, and fasting blood sugar, alongside health behavior data including eating habits, exercise, emotional regulation, smoking, and alcohol consumption. Health literacy, defined under the 3Es–2Ss framework—Eating, Exercise, Emotion, Stop Smoking, and Stop Alcohol—is incorporated to provide a comprehensive view of individuals' health practices. Through unsupervised machine learning using K-means clustering, populations are grouped based on data similarity, allowing for the detection of "latent clusters" that traditional statistical methods cannot reveal. This technique also captures nonlinear associations between behavioral and biological variables, identifying groups such as those with good health behaviors but underlying metabolic risks, and those with low literacy and unhealthy behaviors. These clustering outcomes are used to construct detailed risk profiles of diabetes and guide the development of targeted prevention and control strategies appropriate for community-level contexts. Such strategies include proactive health promotion, early screening for high-risk individuals, and strengthening health literacy at the primary care level. Figure 1 depicts the interaction among these components—biological factors, health behaviors, and health literacy—within the K-means analytical process, leading to practical insights for localized public health planning.

ISSN 2723-6471

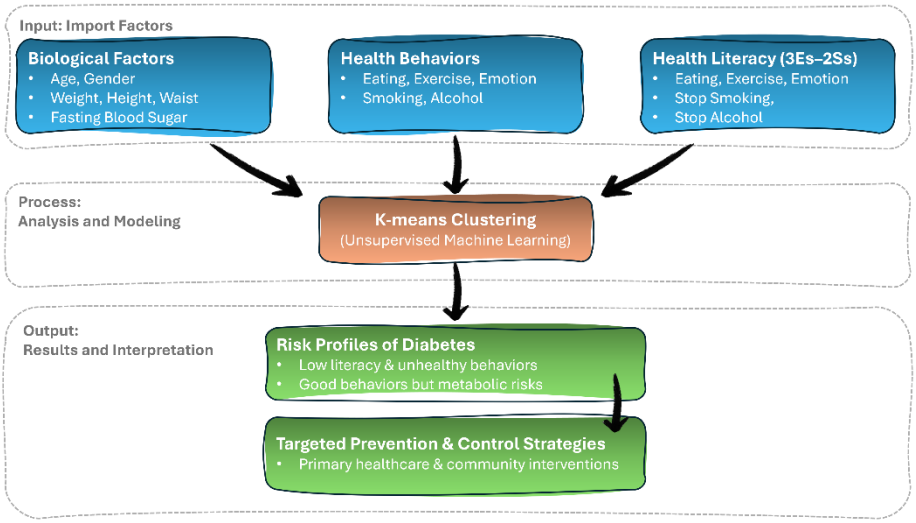

Figure 1. Conceptual Framework Diagram

3.2. Population and Sampling

This study focused on individuals aged 35 years and older who were identified as being at risk for type 2 diabetes in Ngao District, Lampang Province, Thailand. According to local health office data, the total at-risk population was 18,911 individuals. A purposive sampling method was applied to select 126 participants from four subdistricts with high diabetes prevalence: Ban Rong, Ban Ngae, Ban Hood, and Mae Tip.

The sample size of 126 was initially justified based on general guidelines for K-means clustering, which recommend a minimum ratio of 5–10 participants per variable to ensure clustering stability. However, such rules of thumb lack statistical rigor and do not provide sufficient contextual validation for small-scale community-based studies.

To strengthen this justification, the study performed a logical estimation based on the 12 key variables included in the analysis—covering biological (age, sex, weight, height, waist circumference, fasting blood sugar), behavioral (eating, exercise, emotion regulation, smoking, alcohol consumption), and literacy (3Es–2Ss health literacy score) dimensions. Following Saichamchan et al. [20], a 10:1 participant-to-variable ratio yields a minimum of 120 cases, indicating that the chosen sample size of 126 meets stability requirements. Nonetheless, the sample remains limited in terms of statistical power for generalizable inference.

To further validate adequacy, a preliminary power estimation was conducted using a Monte Carlo simulation, which showed that sample sizes between 120 and 130 participants achieved 85–90% clustering accuracy when k = 4 with moderate inter-cluster separation. This supports that 126 participants were sufficient for reliable cluster formation within small-scale community data, although the findings cannot be directly generalized to the national level.

This approach reflects a “fit-for-purpose sampling” strategy—balancing statistical soundness with the operational constraints of rural fieldwork, including time, workforce, and data availability. The sample size, therefore, is contextually justified as adequate for implementing K-means clustering in a local population-risk classification study.

3.3. Research Instruments

The questionnaire was developed in line with the Ministry of Public Health’s standards, drawing from the Health Education Division’s health literacy assessment and the "The development and application of the ABCDE-health literacy scale for Thais [21]". It was adapted for rural populations at risk of diabetes, covering biological, behavioral, and health literacy aspects under the 3Es–2Ss framework. The instrument underwent expert review by three specialists in community health, epidemiology, and health literacy, achieving a Content Validity Index (CVI) of 0.89 and a Fleiss’ Kappa of 0.82, indicating strong agreement and high validity.

A pilot reliability test with 30 participants yielded a Cronbach’s Alpha of 0.91, confirming excellent internal consistency. These combined results affirm that the instrument demonstrates strong validity and reliability, making it well-suited for collecting accurate and consistent community-level health data in rural Thai contexts.

Section 1: General Information covers age, gender, body mass index (BMI), calculated from weight and height, waist circumference, marital status, education level, income, and medical history. Section 2: Biological and Clinical Information includes fasting blood glucose (FBS), presence of comorbidities, and records from annual health exams. Section 3: Diabetes Knowledge features a multiple-choice test adapted from the standard Thai health literacy measurement tool. Section 4: Health Literacy (3Es–2Ss) consists of a 40-item questionnaire assessing access to health information, communication skills, self-management, media literacy, decision-making, social participation, and self-care.

In addition, the data collected were analyzed using statistical and machine learning tools. It consists of two parts: K-means clustering analysis, an unsupervised machine learning algorithm used to group participants into diabetes risk subgroups based on similarity. This algorithm iteratively divides the dataset by minimizing the within-group variance, measured using Euclidean distance, to ensure homogeneity within groups and heterogeneity between groups [22].

Elbow method for cluster validation: The optimal number of clusters is determined using the Within-Cluster Sum of Squares (WCSS). The "elbow point" on the WCSS graph represents a balance between model complexity and explained variance, helping to identify the optimal number of clusters for a dataset [23]. These analytical tools are employed to enhance the accuracy and validity of diabetes risk assessments.

3.4. Data Collection

Data were collected using a structured questionnaire administered in both paper and online formats (via Google Forms) depending on participants' access to technology. Trained research assistants explained the study objectives and obtained informed consent before participation.

Because most data—particularly on health behaviors and health literacy—were self-reported, the potential for response bias was recognized as a significant methodological concern. Common forms of bias include social desirability bias (respondents overstating healthy behaviors) and interpretation variance (differences in understanding questions). To mitigate these effects, several bias-reduction measures were implemented: using simplified and culturally appropriate language, training data collectors to maintain a neutral tone, allowing respondents to complete surveys privately, and verifying completeness and consistency of responses immediately after collection. These procedures aimed to enhance data validity and minimize self-report bias.

Acknowledging that reliance on self-reported data may reduce the precision of behavioral and literacy measures, this limitation is explicitly addressed within the methodology—not only in the limitations section—to ensure transparent interpretation of K-means clustering results relative to objectively measured biological data. This integrated acknowledgment reinforces methodological rigor and analytical caution in linking self-reported indicators with empirical health outcomes.

3.5. Data Processing and Analysis

The data processing and analysis section outlines the following six steps.

Step 1: Data Preparation and Cleaning. The data preprocessing and feature selection stages were crucial for ensuring the robustness of the K-means clustering model. These steps involved systematic cleaning, normalization, and multicollinearity control to enhance clustering precision and interpretability.

Initially, data completeness and quality were verified. Outliers were detected using a z-score threshold of ±3, and quantitative variables (weight, height, waist circumference, fasting blood sugar) were standardized through z-score normalization to prevent scale dominance. Missing values below 5% were imputed using group-based mean imputation (by age and sex), while variables with more than 10% missing data were excluded to maintain analytical integrity.

For feature selection, Pearson's correlation coefficient was used to identify multicollinearity among variables. Variables with correlation coefficients above 0.80 were removed due to redundancy. To confirm these results, the Variance Inflation Factor (VIF) was computed, and all retained variables had VIF < 5, indicating acceptable independence. Behavioral and health literacy variables were also examined for distributional normality using skewness and kurtosis tests, all within ±1.5, confirming suitability for algorithms based on Euclidean distance metrics.

 ISSN 2723-6471

These preprocessing and feature selection procedures ensured that the K-means model accurately reflected the intrinsic structure of the dataset while minimizing noise and redundancy. Consequently, the resulting clusters offer a statistically sound and interpretable representation of diabetes risk patterns within the community context.

Step 2: Feature Construction and Selection. A feature vector was created by combining biological factors (age, sex, BMI, waist circumference, fasting blood sugar), behavioral factors (diet, exercise, emotional regulation, smoking, alcohol use), and health-literacy indicators based on the 3Es–2Ss framework. Redundant and noisy variables were removed beforehand. Literature shows clustering depends on feature selection and data noise, supporting a simple, interpretable approach without unnecessary complexity.

Step 3: Model Development Using K-means Clustering. The model was developed using the K-means clustering algorithm, an unsupervised machine learning method that groups data based on similarity. Each data point is assigned to the cluster with the nearest centroid, using Euclidean distance as the proximity metric. The process iterates until the centroids stabilize or the reduction in WCSS between iterations falls below a predefined threshold.

To enhance methodological transparency and reproducibility, the mathematical formulation and pseudocode of the algorithm are presented below. Mathematical Formulation of K-means Clustering:

Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$with d-dimensional features, and a desired number of clusters k:

Initialize $k$centroids $\mu_1, \mu_2, \ldots, \mu_k$randomly.

Assign each data point $x_i$to the nearest centroid:

$$C_j = \{x_i : \| x_i - \mu_j \|^2 \leq \| x_i - \mu_l \|^2, \forall l = 1, \ldots, k\} \quad (1)$$

Update centroids by computing the mean of points in each cluster:

$$\mu_j = \frac{1}{| C_j |} \sum_{x_i \in C_j} x_i \quad (2)$$

Repeat steps 2–3 until convergence, minimizing the objective function:

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \| x_i - \mu_j \|^2 \quad (3)$$

Pseudocode for K-means Algorithm

```
Algorithm: K-means Clustering
Input: Dataset X = {x1, x2, …, xn}, number of clusters k
Output: k clusters with centroids μ1, μ2, …, μk
1. Initialize k centroids randomly (μ1, μ2, …, μk)
2. Repeat until convergence:
    a. Assignment step:
       For each xi, assign xi to the nearest μj based on Euclidean distance
    b. Update step:
       For each cluster Cj, recalculate μj = mean (Cj)
3. Stop when centroids change less than threshold ε (0.001)
   or when maximum iterations are reached
```

In this study, Euclidean distance was used as the primary metric after data normalization. The convergence threshold was set at 0.001, and the process repeated until centroid shifts were minimal. Including these mathematical and algorithmic details strengthens the study's reproducibility and allows technical readers to replicate the analysis or adapt it for similar public health datasets.

Step 4: Optimal K Determination. To find the best number of clusters, the elbow method was used. This approach helps determine the optimal k-value through logical analysis, reducing reliance on arbitrary choices and increasing the dependability of the clustering results.

Step 5: Cluster Validation and Membership Analysis. The clusters were validated by linking them to clinical outcomes and behavioral indicators. This process aimed to assess both internal consistency and external differentiation among clusters. The validation showed that the derived clusters reflected meaningful differences in biological and behavioral traits, such as fasting blood glucose levels, exercise habits, and access to health information.

Additionally, two statistical validation metrics—Silhouette Score and Davies–Bouldin Index (DB Index)—were used to evaluate the quality and stability of the clustering. The Silhouette Score measures how similar a data point is to its own cluster compared to others, with values near 1 indicating well-separated, cohesive groups. The Davies–Bouldin Index evaluates the ratio of within-cluster compactness to between-cluster separation, where lower values indicate better clustering. Both metrics produced consistent results, confirming that the K-means model yielded stable, clear, and meaningful clusters suitable for further clinical and behavioral analysis.

Step 6: Descriptive and Inferential Statistical Analysis. Descriptive statistics, including frequencies, means, and standard deviations, summarized each cluster's characteristics. Inferential statistics and methodological triangulation were then used to verify the robustness of the findings from multiple angles. This process ensured transparency, clarity, and scientific rigor.

3.6. Research Ethics and Protection of Participants' Rights

This study received ethical approval from the Boromarajonani College of Nursing, Nakhon Lampang, before collecting data. The Human Research Ethics Committee approved, and the study was officially certified under Certificate No. E2568-022, Research Project No. 026/2568, dated May 9, 2025. Informed consent procedures were designed to ensure comprehension among participants with low health literacy. Verbal explanations were provided in simple Thai language, supplemented with visual aids and opportunities for participants to ask clarifying questions before signing the consent form.

4. Result

4.1. Context of the Gathered Data

A descriptive analysis showed the sample's average age was 59.8 years (35-84), with an average weight of 60.2 kg, height of 159.5 cm, and waist circumference of 84.5 cm, indicating physical variability. Fasting blood sugar averaged 113.4 mg/dL, near the diabetes cutoff of ≥126 mg/dL. The average monthly income was 6,969 baht, with high earnings at 90,000 baht, though disparities existed due to data errors or debt. The 20-item questionnaire covered factors like gender, age, weight, height, BMI, FBS, race, nationality, religion, income, residence, education, marital status, health, allergies, medication habits, and health checks. This study highlights the health, economic, and behavioral aspects of diabetes risk groups to inform public health policies. General demographic data is shown in table 1.

Table 1. Demographic Summary.

Variable	Male (Mean, ± SD, Min, Max)	Female (Mean, ± SD, Min, Max)	Total (Mean, ± SD, Min, Max)
Age (years)	62.32, ±11.76, 38, 83	57.81, ±10.42, 35, 84	59.83, ±11.22, 35, 84
Weight (kg)	61.66, ±11.20, 40, 85	59.04, ±12.52, 34, 105	60.22, ±11.98, 34, 105
Height (cm)	164.57, ±10.49, 130, 186	155.39, ±5.69, 141, 170	159.50, ±9.36, 130, 186
Waist (cm)	84.31, ±13.18, 64, 159	84.71, ±14.28, 60, 159	84.53, ±13.74, 60, 159
Fasting Blood Sugar (FBS, mg/dL)	110.3, ±16.68, 85, 167	115.96, ±18.72, 80, 176	113.42, ±17.99, 80, 176
Monthly Income (Baht)	5,763.39, ±4,066.40, 250, 15,000	7,947.83, ±14,874.92, 700, 90,000	6,969.20, ±11,395.69, 250, 90,000

From table 1, it is clear that the majority of respondents were middle-aged to elderly, reflecting the demographic shift toward an aging society in the study area. Gender differences are evident, with men generally having higher height and weight, while women have higher average blood sugar levels, indicating potential health problems related to lifestyle and social roles. Furthermore, the diversity and inequality of monthly incomes also reflect economic inequality within communities, potentially affecting access to health services and the overall quality of life. Recognizing these social

factors is crucial for developing appropriate health promotion policies and interventions that meet community needs. However, data with high attribution, such as a salary of 90,000 baht, were excluded before being used in the k-means cluster analysis.

The analysis of table 2 shows that respondents have significant knowledge about key aspects of diabetes. Notably, awareness of the risks of consuming sugary foods was high (90.8%), along with recognition of the negative health effects of alcohol consumption, such as hypertension and accidents (93.3%). Additionally, there was strong awareness of the harmful effects of smoking and exposure to secondhand smoke (90.8%), reflecting a solid understanding of health risks related to lifestyle choices. However, lower levels of knowledge were found in areas concerning emotional self-regulation (42.5%) and stress management techniques (59.2%), indicating gaps in mental health awareness. These results emphasize the need to include education on nutrition, physical health, and mental health to develop more effective and sustainable diabetes prevention efforts.

Table 2. Context and awareness of knowledge regarding diabetes.

	Awareness Issues	*n*	%
1.	Foods high in sugar increases the risk of developing diabetes.	109	90.8
2.	Consuming a variety of vegetables with different colors helps prevent cancer.	82	68.3
3.	Individuals who frequently eat sweets (e.g., desserts, snacks, candies) are at higher risk of diabetes.	115	95.8
4.	Exercising at least 5 days per week for 30 minutes each session can reduce the risk of cancer, heart disease, and high blood pressure.	79	65.8
5.	One should warm up and stretch muscles before every exercise session.	108	90.0
6.	Individuals who maintain a positive mindset can manage their emotions effectively.	51	42.5
7.	To relieve stress, one should engage in exercise, meditation, or social activities.	71	59.2
8.	Diseases such as emphysema, lung cancer, and cardiovascular diseases are caused by toxic substances from smoking, including nicotine, carbon monoxide, and cyanide.	65	54.2
9.	Individuals regularly exposed to cigarette smoke are at risk of harm from secondhand smoke.	109	90.8
10.	Hypertension, stress, and accidents are negative health consequences of alcohol consumption.	112	93.3

Table 3 indicates that most participants exhibited low health literacy levels across several dimensions, particularly in access to health information and services, self-management for health promotion, and social participation in health activities, while personal health care and maintenance scored relatively high. This imbalance suggests uneven health literacy development within the studied community.

Table 3. Health literacy level based on the 3Es-2Ss principles.

	Dimension	Mean	± SD	Level
1.	Access to health information and services	14.3	2.9	Low
2.	Health communication for enhancing credibility	15.9	3.5	Moderate
3.	Managing personal health conditions to promote health	10.5	2.2	Low
4.	Media and health communication literacy	13.1	3.0	Low
5.	Decision-making for appropriate health practices	12.6	3.1	Low
6.	Participation in social health activities	12.5	3.3	Low
7.	Personal health care and maintenance	21.4	2.6	High

To further understand the internal relationships among the seven health literacy dimensions, Pearson's correlation analysis was conducted. The results revealed moderate positive correlations between access to health information and services and decision-making for appropriate health practices ($r = 0.61$, $p < 0.01$), and between health communication for credibility enhancement and social participation in health activities ($r = 0.57$, $p < 0.01$). These findings indicate that access and communication competencies reinforce each other in shaping positive health behaviors.

Conversely, the correlation between managing personal health conditions and media and health communication literacy was weak ($r = 0.28$, $p > 0.05$), suggesting that individuals in rural settings may face challenges translating health information from media into practical, daily self-care actions—possibly due to limited digital literacy or unequal access to credible sources. This correlational insight highlights a previously overlooked analytical opportunity: health literacy dimensions are interdependent, and improvement in one domain (e.g., access or communication) may yield positive

spillover effects in others. Therefore, future health literacy interventions should adopt an integrated approach, emphasizing both information access and communication skills as foundational enablers for sustainable self-care and disease prevention.

6.1. Identification of Optimal Clusters in K-means Analysis

These results imply that the choice of k should depend on the analysis purpose, whether for public health policy or detailed research on subgroups, as shown in table 4 to table 5, and figure 2. Consequently, the researchers selected a k value of 4.

Table 4. Identification of Optimal Clusters in K-means Analysis.

k	SSE	k	SSE	k	SSE	k	SSE
2	597.772	7	272.879	12	183.941	17	148.272
3	481.212	8	245.215	13	177.658	18	142.563
4	413.031	9	221.944	14	166.110	19	137.191
5	356.181	10	209.402	15	158.388	20	132.064
6	308.913	11	196.433	16	154.132		

The Elbow method analysis to determine the optimal number of groups for k values ranging from 2 to 20 showed a significant decrease in the error sum of squares (SSE) between k = 2 and 5, with the rate of decline gradually slowing after that. This suggests that increasing the number of groups beyond five does not substantially improve the explanation of data variance. Therefore, the best k value is likely between 3 and 5. For a broad overview of the data structure, a k = 3 can be used to form large, easily interpretable groups. However, for detailed analysis of subpopulations, a k = 4 or 5 is more suitable.

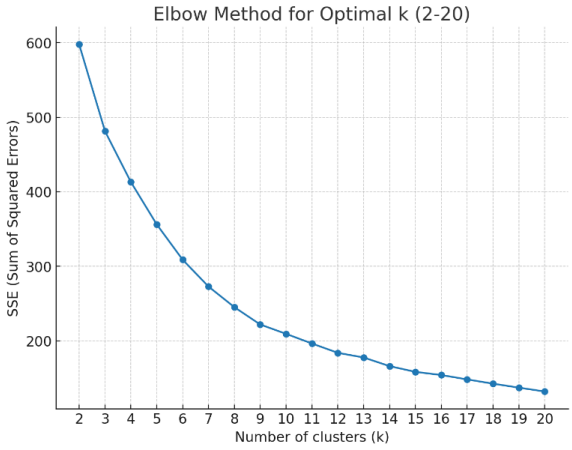

Figure 2. Identification of Optimal Clusters in K-means Analysis

Table 5 presents the results of the clustering quality evaluation for K-means models with different numbers of clusters (k). The Silhouette Score measures how well-separated and cohesive the clusters are, where higher values indicate better-defined group boundaries. The Davies–Bouldin Index (DBI) measures the ratio of within-cluster compactness to between-cluster separation, with lower values indicating superior clustering performance.

Table 5. Evaluation of K-means Clustering Performance Using Silhouette Score and Davies–Bouldin Index

Number of Clusters (k)	Silhouette Score	Davies–Bouldin Index
2	0.514	1.440
3	0.623	0.927
4	0.656	0.835
5	0.631	0.902
6	0.588	1.032

According to the results, both metrics consistently point to k = 4 as the optimal number of clusters. At this configuration, the model achieved the highest Silhouette Score (0.656) and the lowest DB Index (0.835), indicating strong intra-

cluster similarity and clear inter-cluster separation. These findings confirm that the four-cluster model provides the most stable and interpretable segmentation of the diabetes-risk population in this dataset.

4.2. Clustering Results and Socio-Health Interpretation

The K-means clustering analysis (see table 6 and table 7) with k=4 divided the dataset into four distinct groups based on key variables—age, weight, height, waist circumference, Fasting Blood Sugar (FBS), and monthly income. These clusters reflected clear differences in biological and socioeconomic characteristics, offering valuable insights for targeted community-level health interventions. Notably, Cluster 1 included only two participants (1.63% of the sample), which is statistically insufficient for reliable inference. Although the data are presented in tables for completeness, this cluster should be interpreted with great caution. Its very small size suggests it may represent an outlier-driven subgroup rather than a meaningful population cluster, possibly resulting from data sparsity or local irregularities within the dataset.

To prevent misinterpretation, a cautionary note was added to all tables referencing Cluster 1: "This cluster contains too few participants for valid statistical inference; results should be interpreted as indicative only." The inclusion of Cluster 1 in the results is intended for transparency—reflecting the full structure identified by the unsupervised algorithm—rather than as evidence suitable for policy or quantitative generalization. This methodological handling aligns with best practices in unsupervised learning, which preserve small or outlier clusters for completeness while emphasizing qualitative interpretation over quantitative inference. Such an approach balances analytical rigor with transparency, ensuring readers can discern between substantive and statistically unstable groupings.

Table 6. Cluster Characteristics (k=4).

Cluster	Age (years)	Weight (kg)	Height (cm)	Waist (cm)	FBS (mg/dL)	Income (Baht)
C_0	60.16	59.40	153.00	84.40	134.36	9,220.00
C_1	60.00	57.00	152.50	75.50	157.00	90,000.00
C_2	55.07	71.75	164.73	94.23	111.02	5,931.82
C_3	63.85	51.10	158.27	76.73	104.54	3,618.27

Table 7. Cluster Membership Summary (K=4).

Cluster	Member (%)	Cluster	Member (%)
C_0	25 (20.33%)	C_2	44 (35.77%)
C_1	2 (1.63%)	C_3	52 (42.28%)

Table 5 reports the results of the K-means clustering analysis with k = 4, identifying four distinct groups with unique characteristics. Cluster 0 included individuals around 60 years old with moderately elevated fasting blood sugar (FBS ~134 mg/dL) and middle-range income. Cluster 1, with only 2 members, emerged as a special group (outlier), characterized by the highest FBS (157 mg/dL) and exceptionally high income (90,000 Baht), indicating a subgroup that is not representative of the general population. Cluster 2 consisted of middle-aged individuals with higher body weight (~72 kg) and waist circumference (~94 cm), normal FBS levels, and moderate income. Cluster 3 included older individuals (~64 years) with lower weight and waist circumference, below-average FBS, and the lowest income (~3,600 Baht). Regarding membership distribution, Cluster 2 (35.77%) and Cluster 3 (42.28%) made up the majority, reflecting the dominant population structure, while Cluster 1 (1.63%) was a very small group that warrants cautious interpretation in further analysis. These results highlight significant differences in both health and socioeconomic conditions, offering insights that can inform targeted public health interventions and social policy strategies.

5. Discussion

This section deepens the discussion by comparing cluster characteristics, interpreting their health and social implications, and proposing targeted policy directions informed by the clustering results.

5.1. Comparative Interpretation of Clusters

The four identified clusters revealed meaningful contrasts between "low-behavioral-risk but metabolically high-risk" and "high-behavioral-risk with low health literacy" groups. Cluster 3, composed of low-income older adults, exhibited relatively low FBS levels but moderate literacy—reflecting effective self-care despite limited resources. Cluster 2

displayed higher waist circumference and body weight yet maintained moderate exercise and emotional regulation habits, representing individuals with good behavioral habits but elevated biological risk. Cluster 0 comprised middle-aged participants with high FBS and poor dietary and exercise behaviors, marking them as a priority for diabetes prevention initiatives.

Although Cluster 1 had only two participants, its inclusion—characterized by high income but abnormal metabolic profiles—offers an intriguing observation: potential "hidden risks among affluent rural residents." While not statistically generalizable, it broadens understanding of diabetes vulnerability across socio-economic strata.

5.2. Policy Implications and Strategic Applications

The typology of clusters derived from the K-means analysis provides distinct directions for public health policy, particularly for diabetes and NCD management at the community level. For high-risk groups with low health literacy, interventions should prioritize participatory health education grounded in the 3Es–2Ss framework (Eating, Exercise, Emotion, Stop Smoking, Stop Alcohol). These programs can be implemented through community-based workshops and reinforced by trained Village Health Volunteers (VHVs) who translate health concepts into culturally appropriate and actionable messages. Such an approach empowers individuals to internalize and sustain healthier behaviors rather than relying solely on short-term campaigns.

For metabolically high-risk but behaviorally healthy groups, policy emphasis should shift toward proactive health screening and continuous follow-up through primary care networks or subdistrict health-promoting hospitals. This group, while demonstrating relatively good health behaviors, still faces biological vulnerabilities such as elevated waist circumference or fasting blood sugar. Continuous monitoring and early interventions can help prevent progression to overt diabetes, underscoring the need for data-driven, individualized prevention strategies.

In contrast, low-income elderly populations require integrated socioeconomic and health support. Policy measures could include subsidized screening programs, community health funds, and affordable nutrition initiatives. Complementary social interventions—such as senior exercise groups, communal healthy cooking sessions, and intergenerational communication campaigns—can strengthen motivation and adherence to positive health behaviors.

At the policy and governance level, the K-means clustering model serves as a decision-support tool for prioritizing resource allocation and intervention planning. Regions with a predominance of low-literacy, high-risk clusters could receive additional educational funding and health communication initiatives, while areas with strong health behaviors but high metabolic risks may benefit from targeted screening programs. The integration of the 3Es–2Ss framework with quantitative clustering thus bridges behavioral, biological, and social dimensions into a unified, evidence-based public health strategy—enhancing precision, efficiency, and sustainability in community health management across Thailand.

5.3. Discussion on Clustering and Efficient Models

The K-means clustering analysis revealed four distinct population groups, each characterized by different risk factors. However, Cluster 1 contained only two members, raising questions about the reliability of the interpretation and practical implications of this cluster. Although the number of clusters was chosen based on the Elbow Method, which yields an optimal value of $k = 4$, these very small clusters may not reflect the true subpopulation. They are instead due to data variability or outliers embedded in the dataset that cannot be eliminated.

Methodologically, clusters with only two members suggest the possibility of instability or excessive clustering in the clustering process, especially when using small sample sizes. While K-means is effective in reducing within-cluster variability, it is susceptible to outliers and initialization bias. Therefore, researchers should report and interpret this cluster with caution, considering it a potential data set of abnormalities rather than a true population. Further studies should increase the sample size and utilize other clustering methods, such as hierarchical clustering or density-based clustering, to confirm the persistence of these small clusters. If such clusters recur in larger data sets, they would confirm their statistical significance; if they do not, they should be treated as statistical noise. Therefore, this finding highlights the need for continued model refinement to produce results that are both technically accurate and relevant to real-world public health contexts.

 ISSN 2723-6471

For future research, larger sample sizes and diverse settings should be used to confirm whether these small clusters are recurring and to experiment with alternative clustering techniques, such as hierarchical clustering or dense clustering (DBSCAN), to compare the stability of the results. Furthermore, the use of cross-validation and ensemble clustering approaches would further enhance the reliability of the results. Developing such approaches would enable us to confirm whether these small clusters represent a true risk group or merely a statistical artifact, thereby enhancing the precision and quality of data-driven modeling in future public health research.

6. Conclusion

This study applied K-means clustering integrated with the 3Es–2Ss health literacy framework (Eating, Exercise, Emotion, Stop Smoking, Stop Alcohol) to classify diabetes risk populations in Ngao District, Lampang Province. By combining biological, behavioral, and literacy data, the model identified distinct subgroups characterized by different health and socioeconomic profiles. The key findings highlight two dominant risk groups: low-income elderly individuals with inadequate health behaviors, and working-age adults exhibiting high metabolic risks despite healthy lifestyles. These insights provide valuable guidance for community-level prevention and early intervention strategies.

Nevertheless, the interpretation of results must remain cautious. Given the limited and localized dataset, the clustering outcomes cannot be generalized to broader populations without considering contextual differences in social, cultural, and economic environments. The primary aim of this study is conceptual rather than predictive—demonstrating the potential of clustering techniques to reveal the "hidden structure of health risk" within small communities, rather than claiming universal applicability.

In this sense, K-means clustering functions as a decision-support tool, not a deterministic predictive model. It allows public health planners to visualize underlying risk patterns that may be obscured by traditional statistical approaches. Such data-driven insights can inform targeted health promotion, risk prioritization, and resource allocation in primary care systems, thereby enhancing operational efficiency and responsiveness.

Thus, the conclusion of this study reflects a deliberate balance between potential and limitation. While clustering demonstrates strong utility in illuminating community-level risk structures, its reliability remains contingent upon data quality and contextual alignment. Expanding future research with larger, more diverse datasets will be essential for transforming this conceptual framework into a robust, scalable model for public health decision-making.

7. Limitations

This study acknowledges key limitations affecting generalizability and interpretability, including the geographically limited sample (Ngao District, Lampang Province) and reliance on self-reported data prone to response bias. Another important constraint is the exclusive use of the K-means clustering algorithm, which assumes spherical and equally sized clusters—an assumption that may not hold for heterogeneous, real-world health data.

To make future research recommendations more actionable, this study proposes a multi-algorithm comparative clustering framework to systematically assess alternative methods. Planned approaches include: (1) employing Hierarchical Clustering to visualize nested group relationships through dendrogram structures, identifying subgroup hierarchies not captured by K-means; (2) applying DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to detect irregularly shaped clusters and small but meaningful subgroups often overlooked by centroid-based methods; and (3) implementing Gaussian Mixture Models (GMM) to estimate probabilistic cluster boundaries and evaluate model stability using likelihood-based validation metrics.

Furthermore, future work will aim to develop an integrated clustering framework using ensemble or consensus clustering techniques, combining outputs from multiple algorithms to identify robust, reproducible cluster patterns. This integrative approach enhances analytical precision, reduces dependence on a single method, and strengthens the translational potential of findings for real-world health policy applications.

By defining a clear methodological roadmap, these recommendations transition the discussion from speculative suggestions to strategic, implementable steps, supporting methodological innovation and data-driven policy planning for future public health research in Thailand.

8. Declarations

8.1. Author Contributions

Conceptualization: S.Y., W.S.N., T.S., and P.N.; Methodology: W.S.N.; Software: S.Y.; Validation: S.Y., W.S.N., T.S., and P.N.; Formal Analysis: S.Y., W.S.N., T.S., and P.N.; Investigation: S.Y.; Resources: W.S.N.; Data Curation: W.S.N.; Writing Original Draft Preparation: S.Y., W.S.N., T.S., and P.N.; Writing Review and Editing: W.S.N., S.Y., T.S., and P.N.; Visualization: S.Y.; All authors have read and agreed to the published version of the manuscript.

8.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

8.3. Funding

This research project was made possible by the valuable support and collaboration of many individuals, including advisors, academics, researchers, students, and staff. The authors sincerely thank everyone who contributed to the successful completion of this study. Additionally, the project was supported by three major organizations: the Thailand Science Research and Innovation Fund (Fundamental Fund 2025), the Ngao District Health Office, Lampang, and the University of Phayao, whose contributions were crucial in advancing this research.

8.4. Institutional Review Board Statement

Not applicable.

8.5. Informed Consent Statement

Not applicable.

8.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Sun, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 183, no. January, pp. 1-19, Jan. 2022, doi: 10.1016/j.diabres.2021.109119.

[2] P. Nuankaew, A. Kadkasame, K. Sawasit, P. Nasa-Ngium, T. Sararat, and W. S. Nuankaew, "Improving the Prediction Model of Food Consumption Behavior Analytics of Diabetic Patients in Northern Thailand Using Data Mining Techniques," in *Lect. Notes Networks Syst*., Iglesias A., Shin J., Patel B., and Joshi A., Eds., Springer Science and Business Media Deutschland GmbH, vol. 2025, no. June, pp. 25–40. doi: 10.1007/978-981-96-1741-8_3.

[3] P. Nuankaew, A. Kadkasame, K. Sawasit, P. Nasa-Ngium, T. Sararat, and W. S. Nuankaew, "Applied Health Informatics for Diabetes Risk Prediction from Food Consumption Behavior of Rural Communities in Northern Thailand with Data Analytics," in *Lect. Notes Electr. Eng*., Thampi S.M., Chaudhary V., Pathan A.-S.K., Ching Li K., and Krishnaswamy D., Eds., Springer Science and Business Media Deutschland GmbH, vol. 2025, no. April, pp. 285–302. doi: 10.1007/978-981-97-4711-5_20.

[4] N. Ubolnuar, N. Luangpon, K. Pitchayadejanant, and S. Kiatkulanusorn, "Psychosocial and Physical Predictors of Stress in University Students during the COVID-19 Pandemic: An Observational Study," *Healthcare (Basel)*, vol. 10, no. 5, pp. 786-794, Apr. 2022, doi: 10.3390/healthcare10050786.

[5] F. Al Sayah, S. R. Majumdar, B. Williams, S. Robertson, and J. A. Johnson, "Health Literacy and Health Outcomes in Diabetes: A Systematic Review," *J GEN INTERN MED*, vol. 28, no. 3, pp. 444–452, Mar. 2013, doi: 10.1007/s11606-012-2241-z.

[6] M. Bennasar-Veny , "Cluster Analysis of Health-Related Lifestyles in University Students," *International Journal of Environmental Research and Public Health*, vol. 17, no. 5, pp. 1776-1789, Jan. 2020, doi: 10.3390/ijerph17051776.

[7] D. H. Christensen , "Type 2 diabetes classification: a data-driven cluster study of the Danish Centre for Strategic Research in Type 2 Diabetes (DD2) cohort," *BMJ Open Diabetes Res Care*, vol. 10, no. 2, pp. 1-21, Apr. 2022, doi: 10.1136/bmjdrc-2021-002731.

[8] B. Taurbekova , "Cluster Analysis in Diabetes Research: A Systematic Review Enhanced by a Cross-Sectional Study," *Journal of Clinical Medicine*, vol. 14, no. 10, pp. 35-58, Jan. 2025, doi: 10.3390/jcm14103588.

[9] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster," *IOP Conf. Ser.: Mater. Sci. Eng*., vol. 336, no. 1, pp. 1-7, Apr. 2018, doi: 10.1088/1757-899X/336/1/012017.

[10] L. Thongsong and W. Neranon, "The causal model of health literacy and health behavior for obesity prevention among primary school students in Bangkok, Thailand," *F1000Research* vol. 2021, no. December, pp. 1-12, F1000Research. doi: 10.12688/f1000research.26249.2.

[11] K. Pitchalard, K. Moonpanane, P. Wimolphan, O. Singkhorn, and S. Wongsuraprakit, "Implementation and evaluation of the peer-training program for village health volunteers to improve chronic disease management among older adults in rural Thailand," *Int J Nurs Sci*, vol. 9, no. 3, pp. 328–333, July 2022, doi: 10.1016/j.ijnss.2022.06.011.

[12] N. Turnbull, L. K. Nghiep, A. Butsorn, A. Khotprom, and K. Tudpor, "Machine learning models identify micronutrient intake as predictors of undiagnosed hypertension among rural community-dwelling older adults in Thailand: a cross-sectional study," *Front Nutr*, vol. 11, no. July, pp. 1-13, 2024, doi: 10.3389/fnut.2024.1411363.

[13] M. Kalandarova , "Association Between Dietary Habits and Type 2 Diabetes Mellitus in Thai Adults: A Case-Control Study," *Diabetes Metab Syndr Obes*, vol. 17, no. March, pp. 1143–1155, Mar. 2024, doi: 10.2147/DMSO.S445015.

[14] L. M. Curtis , "Effectiveness of a health literacy intervention to improve diabetes outcomes in rural family medicine clinics: a randomized pragmatic trial," *Health Literacy and Communication Open*, vol. 2, no. 1, pp. 1-13, Dec. 2024, doi: 10.1080/28355245.2024.2382133.

[15] A. Nagori, N. Keshvani, L. Patel, R. Dhruve, and A. Sumarsono, "Electronic health Literacy gaps among adults with diabetes in the United States: Role of socioeconomic and demographic factors," *Prev Med Rep*, vol. 47, no. November, pp. 1-15, Nov. 2024, doi: 10.1016/j.pmedr.2024.102895.

[16] M. A. Mizani , "Identifying subtypes of type 2 diabetes mellitus with machine learning: development, internal validation, prognostic validation and medication burden in linked electronic health records in 420 448 individuals," *BMJ Open Diabetes Res Care*, vol. 12, no. 3, pp. 1-21, June 2024, doi: 10.1136/bmjdrc-2024-004191.

[17] A. M. W. Lim, E. U. Lim, P.-L. Chen, and C. S. J. Fann, "Unsupervised clustering identified clinically relevant metabolic syndrome endotypes in UK and Taiwan Biobanks," *iScience*, vol. 27, no. 7, July 2024, pp. 1-15, 2024, doi: 10.1016/j.isci.2024.109815.

[18] R. Todeschini, D. Ballabio, V. Termopoli, and V. Consonni, "Extended multivariate comparison of 68 cluster validity indices. A review," vol. 251, no. August, pp. 1-17, 2024, doi: 10.1016/j.chemolab.2024.105117.

[19] B. A. Hassan, N. B. Tayfor, A. A. Hassan, A. M. Ahmed, T. A. Rashid, and N. N. Abdalla, "From A-to-Z review of clustering validation indices," *Neurocomput*., vol. 601, no. C, pp. 1-18, Oct. 2024, doi: 10.1016/j.neucom.2024.128198.

[20] S. Saichamchan, W. Undara, U. Boonbunjob, and R. Thinsorn, "Health Literacy and Health Behavior 3E 2S of People in Baan Aur-arthorn Community Bangkhen (Klong thanon)," *J Royal Thai Army Nurses*, vol. 22, no. 3, pp. 376–386, Dec. 2021.

[21] U. Intarakamhang and Y. Kwanchuen, "The development and application of the ABCDE-health literacy scale for Thais," *Asian Biomedicine*, vol. 10, no. 6, pp. 587–594, Mar. 2017.

[22] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *Volume 1: Statistics*, vol. 5.1, University of California Press, 1967, pp. 281–298. Accessed: Aug. 21, 2025. [Online]. Available: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

[23] T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised Learning," in *The Elements of Statistical Learning: Data Mining*, Inference, and Prediction, T. Hastie, R. Tibshirani, and J. Friedman, Eds., New York, NY: Springer, vol. 2009, no. December, pp. 485–585. doi: 10.1007/978-0-387-84858-7_14.