AN ADOPTION OF A STACKING ENSEMBLE METHOD TO PREDICT NONCOMMUNICABLE DISEASES IN THAILAND

AN ADOPTION OF A STACKING ENSEMBLE METHOD TO PREDICT NONCOMMUNICABLE DISEASES IN THAILAND

A thesis Presented to
The Graduate School of Bangkok University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Information Technology and Data Science

by
Peat Winch
2024

This thesis has been approved by
the Graduate School
Bangkok University

Title: An Adoption of a Stacking Ensemble Method to Predict Noncommunicable Diseases in Thailand

Author: Peat Winch

Thesis Committee:

Chairman (External Representative)	Asst. Prof.Dr.Chorkaew Jaturanonda
Committee (Thesis Advisor)	Dr.Nattapong Sanchan
Committee (Thesis Co-advisor)	Assoc.Prof.Anon Sukstrienwong
Committee (Program Faculty Members)	Asst.Prof.Dr.Patravadee Vongsumedh

Winch, Peat., PharmD, Master of Science, Information Technology and Data Science), March 2025, Graduate School, Bangkok University
An Adoption of a stacking ensemble method to predict noncommunicable diseases in Thailand **(**83 pp**.)**
Advisor of dissertation: Nattapong Sanchan, Ph.D.

ABSTRACT

This study delved into predictive modeling for Non-Communicable Diseases (NCD) prevalence in Thailand, focusing on the significance of Social Determinants of Health (SDH) related features. Through an extensive analysis of various datasets and machine learning models, including Support Vector Regression (SVR), Gradient Boosting Decision Tree (GBDT), Linear Regression (LR), Random Forest (RF), Stacking, and XGBoost, the research evaluated predictive capabilities and explanatory power across different scenarios.

Fin dings revealed the importance of socioeconomic and environmental factors such as household income, air pollution levels, education-related variables, household expenses, and healthcare in predicting NCD occurrence or progression. While SVR occasionally exhibited lower Mean Absolute Error (MAE), it struggled with poor explanatory power, as evidenced by negative or low R-squared and Adjusted R-squared values. Other models, particularly GBDT, RF, and XGBoost, consistently demonstrated superior predictive accuracy and moderate to better explanatory capabilities across various scenarios.

The study highlighted challenges including dataset discrepancies, lack of data granularity, and the need for more detailed features, urging future research to address these limitations. Further exploration of additional SDH, incorporation of advanced machine learning techniques, longitudinal studies, and expansion of datasets to include larger and more diverse populations were suggested for improving predictive models' accuracy and explanatory power. These insights offered valuable guidance for healthcare practitioners and policymakers in devising evidence-based strategies to mitigate NCD's impact on public health.

Keywords: Noncommunicable Diseases, Social Determinants of Health, Stacking Ensemble Method, Prediction

TABLE OF CONTENTS

	Page
ABSTRACT……..……	iii
LIST OF TABLES……..……..	viii
LIST OF FIGURES……..	ix
CHAPTER 1: INTRODUCTION…….	1
1.1 Background……..	1
1.2 Research Question……...	2
1.3 Research Objective……..	4
1.4 Conceptual Framework……	5
1.5 Scope of Research……...	6
CHAPTER 2: LITERATURE REVIEW……..	8
2.1 The Relation of SDH and NCD……..	8
2.2 Existing Predictive Algorithms for NCD……	9
2.2.1 Supervised Methods……...	12
2.2.2 Unsupervised Methods……...	13
2.2.3 Other Methods……...	13
2.3 SDH-Related Features in Previous Works……..	20
2.4 Rationale for Adopting the Stacking Ensemble Methodology…...	23
CHAPTER 3: METHODOLOGY……	25
3.1 Software Environment…….	25
3.2 Data Collection Framework……	25
3.2.1 Strategic Approach…….	25
3.2.2 Data Source Overview……...	26
3.3 Dataset Characteristics……	30
3.3.1 NCD Prevalence Dataset Overview……...	32
3.3.2 SDH Dataset Overview……..	33
3.3.3 Trends Analysis……..	34

TABLE OF CONTENTS (Continued)

Page

CHAPTER 3: METHODOLOGY (Continued)…… 39
3.4 Data Pre-Processing…… 39
3.4.1 Data Cleaning and Manipulation…… 39
3.4.2 Missing Value…… 41
3.4.3 Outlier……. 41
3.5 Model Adoption…….. 42
3.5.1 Algorithm Selection and Implementation…….. 43
3.6 Validation and Performance Assessment…… 44
3.6.1 Cross Validation Strategy……... 44
3.6.2 Performance Metrics…….. 45
CHAPTER 4: FINDINGS……. 47
4.1 Feature Importance Analysis……... 47
4.1.1 Feature Importance in CA…….. 49
4.1.2 Feature Importance in CVD……... 49
4.1.3 Feature Importance in DM……. 50
4.1.4 Feature Importance in HTN……... 51
4.1.5 Feature Importance in COPD……. 51
4.1.6 Feature Importance in Stroke……. 51
4.2 Model Evaluation…… 55
4.2.1 The Result Description for Baseline Scenario……... 55
4.2.2 The Result Description for Inference Scenario…….. 55
4.2.3 Findings Summary……. 58
CHAPTER 5: DISCUSSION…… 61
5.1 Conclusion……... 61
5.2 Discussion……... 61
5.3 Limitation…… 64
5.4 Future Work…… 65

TABLE OF CONTENTS (Continued)

Page

BIBLIOGRAPHY……. 68

APPENDIX……... 74

Appendix 1: Publication…… 75

Appendix 2: Abbreviations……... 76

BIODATA……. 83

LIST OF TABLES

		page
Table 2.1:	Algorithms Used by Category…….	12
Table 2.2:	Forecast Model Review: Individual Attributes……...	14
Table 2.3:	Forecast Models: Review of Non-Individual Attributes…….	18
Table 2.4:	Previous Studies' Attribute Coverage Across SDH Domains….	22
Table 3.1:	Data Sources and Strategic Approach…….	27
Table 3.2:	Overview of Data Characteristics……...	31
Table 3.3:	Top 10 Highest Patients (2012 – 2021)……..	32
Table 3.4:	Dataset Feature: SDH Feature…….	33
Table 3.5:	Minimum, Maximum, Average, Medium, and Mode of NCD Patients……	41
Table 4.1:	Feature Importance Score……...	53
Table 4.2:	Summary of Comparison of Model Performance Metrics Across Different NCDs in Baseline and Inference Scenarios….	56

LIST OF FIGURES

page

Figure 2.1: Literature Review Consort…… 10

Figure 2.2: Feature Analysis & Deployment Rates (16 Studies)……. 11

Figure 3.1: Analyzing the Evolution of NCD Patient Trends in Thailand…….. 36

Figure 3.2: Spatial Analysis of Cumulative NCD Rates in Thailand…….. 37

Figure 3.3: Geographic Patterns of SDH in Thailand…….. 38

Figure 3.4: Historical Trends of Household Financial Profiles in Thailand…… 39

Figure 3.5: Diagram of Modelling……... 43

Figure 4.1: Feature Importance Score in Visualisation……... 48

CHAPTER 1
INTRODUCTION

1.1 Background

World Health Organization (2022) described Noncommunicable diseases (NCD) that "NCD are tended to be of long duration and the result of a combination of genetic, physiological, environmental and behavioural factors". The risk factors of NCD include modifiable behavioural and metabolic risk factors; smoking accounts is one of the most impact modifiable behavioural risk factors due to its death rate of over 8 million a year. The consequence of NCD affecting society includes the cost of illness contributed to NCD, productivity loss, and a substantial hidden cost affecting public policy planning and the loss of labour driving the economy in that particular community.

The selection of specific NCDs for this study including Diabetes Mellitus (DM), Hypertension (HTN), Chronic Obstructive Pulmonary Disease (COPD), Cardiovascular Disease (CVD), Stroke, and Cancer (CA), was guided by their substantial burden on Thailand's healthcare system and their demonstrated sensitivity to socioeconomic and environmental factors. These six conditions represent the leading causes of mortality and morbidity in Thailand, collectively accounting for more than 70% of all deaths nationwide (Ministry of Public Health of Thailand et al., 2021). DM affects approximately 8.9% of the Thai adult population, while HTN prevalence reaches 24.7%, with both conditions showing marked disparities across socioeconomic strata (Nawamawat et al., 2020). COPD, with a prevalence of 6.8% among Thai adults over 40, demonstrates strong associations with environmental exposures and smoking behaviors that vary significantly by region and socioeconomic status (Potempa et al., 2022). Similarly, CVD and Stroke account for 23% of all deaths in Thailand, with incidence patterns closely mirroring geographic variations in healthcare access and economic development (World Health Organization, 2022). Cancer mortality rates in Thailand have increased by 30% over the past decade, with pronounced disparities between urban and rural populations suggesting significant influence from social determinants (Ministry of Public Health of Thailand et al.,

2021). The selection of these specific conditions provides a comprehensive framework for evaluating how Social Determinants of Health (SDH) impact the most significant NCD burdens in Thailand, offering potential insights for targeted public health interventions across diverse disease categories.

A problem in the field of NCD prevalence prediction is the persistent reliance on individual medical factors rather than population-level social determinants (Bhoothookngoen & Sanchan, 2023). This narrow focus has created a significant knowledge gap regarding how broader social, economic, and environmental conditions affect disease patterns across populations. Despite the WHO's recognition of SDH as crucial health influencers (World Health Organization, 2022), most predictive models continue to overlook these factors, particularly in middle-income countries like Thailand (Nawamawat et al., 2019). The fragmented approach to studying SDH domains in isolation rather than as an interconnected system further compounds this problem (Stringhini et al., 2018; Wang & Wang, 2020). This limitation restricts our understanding of how these domains collectively shape disease patterns and hinders the development of effective, holistic public health strategies (Potempa et al., 2022). Additionally, methodological challenges in handling inconsistent SDH datasets with missing values across different provinces and time periods have further complicated research efforts, leading to potentially biased or unreliable predictions (Hu et al., 2020). These problematic gaps directly informed the research questions of this study, which sought to address how effectively stacking ensemble methods utilizing SDH features could predict NCD prevalence, which specific SDH features demonstrated the highest predictive importance for different NCD categories, and how different data preprocessing strategies affected model performance when working with Thailand's spatially inconsistent datasets.

1.2 Research Question

1.2.1 How effectively can a stacking ensemble methodology utilizing SDH (population-level) features predict NCD prevalence across Thailand? This research question examined the application of a stacking ensemble methodology for forecasting NCD prevalence using population-level Social Determinants of Health. The stacking approach, as implemented

by Hu et al. (2020), combined multiple algorithms to capture complex relationships within heterogeneous data. This method was particularly valuable for analyzing the multidimensional SDH factors incorporated in this research: economic indicators (household income, expenses, and loans), educational metrics (years of schooling, number of educational institutions), healthcare infrastructure (hospital distribution), environmental conditions (pm2.5 levels), and social context variables (smoking rates, alcohol consumption). By evaluating how effectively these population-level factors predicted disease patterns, this study aimed to shift the focus from individual clinical risk factors towards broader social determinants that could inform policy-level interventions (Bhoothookngoen & Sanchan, 2024).

1.2.2 Which specific SDH features demonstrate the highest predictive importance for different NCD categories (DM, HTN, COPD, CVD, Stroke, and Cancer) in Thailand?

This research question sought to identify the most influential social determinants for each NCD category examined. Understanding which specific SDH features, whether economic, educational, healthcare-related, environmental, or social, contributed most significantly to different disease patterns provided valuable guidance for targeted public health interventions. For instance, if air pollution (pm2.5) demonstrated high importance for respiratory conditions like COPD, while economic factors showed stronger associations with diabetes, policymakers could prioritize different intervention strategies for different disease categories. This question aimed to create an evidence-based hierarchy of social determinants for each condition, enabling more efficient resource allocation and more targeted preventive measures across Thailand's diverse provinces (World Health Organization, 2022; Potempa et al., 2022).

1.2.3 How to evaluate the NCD prediction?

This research question addressed the methodological approaches for assessing prediction quality when working with Thailand's spatially

inconsistent health data. The study employed multiple complementary performance metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) to evaluate prediction accuracy, alongside R^2 and Adjusted R^2 to assess explanatory power. These metrics were applied across two distinct scenarios: a baseline scenario using complete case analysis (removing records with missing values) and an inference scenario employing mean imputation (replacing missing values with feature averages). This dual approach allowed for systematic comparison of how different data preprocessing strategies affected model performance and reliability, providing methodological guidance for future health forecasting efforts using similar datasets (Hu et al., 2020; Stringhini et al., 2018).

1.3 Research Objective

The objectives of this research have been focused on aligning with the research question:

1.3.1 To apply a stacking ensemble methodology proposed by Hu et al.'s (2020) approach to evaluate SDH features' predictive capabilities for NCD prevalence across Thailand.

1.3.2 To identify the significant SDH features for each NCD category, creating an evidence-based hierarchy to guide targeted public health interventions.

1.3.3 To compare data preprocessing strategies, specifically complete case analysisversus imputation methods, when working with spatially inconsistent SDH datasets. This comparison will employ a comprehensive evaluation framework utilizing multiple performance metrics (MAE, RMSE, MAPE, R^2, and Adjusted R^2) to assess both prediction accuracy and variance explanation capabilities across different NCD categories, thereby providing quantifiable evidence for determining optimal preprocessing approaches for Thailand's provincial health data.

1.4 Conceptual Framework

This thesis was guided by an integrated conceptual framework that connected SDH with NCD prevalence through a stacking ensemble method. The framework consisted of three interconnected components that collectively addressed the research questions and objectives.

The first component encompassed the SDH Features, organized according to the five domains established by ODPHP: economic stability (household income, expenses, and loans), education access and quality (years of schooling, educational institutions), healthcare access and quality (hospital distribution), neighborhood and built environment (pm2.5 levels), and social and community context (smoking rates, alcohol consumption). These population-level indicators represented the complex social fabric that influenced health outcomes across Thailand.

The second component focused on NCD Prevalence Patterns across six conditions of significant public health concern in Thailand: DM, HTN, COPD, CVD, Stroke, and CA. These conditions were selected based on their substantial burden on Thailand's healthcare system (Ministry of Public Health of Thailand et al., 2021; Potempa et al., 2022) and their established sensitivity to socioeconomic and environmental factors (Nawamawat et al., 2020; World Health Organization, 2022).

The third component consisted of the Methodological Framework, centered on a stacking ensemble approach adapted from Hu et al. (2020). This approach incorporated multiple base algorithms (Linear Regression (LR), Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Gradient Boosting Decision Tree (GBDT)) in the first stage, with RF serving as the meta-learner in the second stage. This component also included two distinct data preprocessing strategies, complete case analysis (baseline scenario) and mean imputation (inference scenario), allowing for systematic comparison of how missing value treatment affected model performance.

The integration of these three components created a comprehensive framework for investigating how different configurations of social determinants predicted disease patterns, which features exerted the strongest influence on specific

conditions, and how methodological choices impacted predictive accuracy. This framework addressed the specific research questions by:

1.4.1 Enabling systematic evaluation of how effectively stacking ensemble methods utilizing SDH features could predict NCD prevalence at the population level across Thailand's provinces

1.4.2 Incorporating feature importance analysis to identify which SDH features contributed most significantly to predictive accuracy for each NCD category

1.4.3 Facilitating direct comparison between complete case analysis and mean imputation strategies to determine optimal approaches for handling missing values in spatially inconsistent datasets

The findings derived from this framework provided evidence-based insights to guide targeted public health interventions and resource allocation, ultimately contributing to more effective NCD prevention and management strategies in Thailand.

1.5 Scope of Research

This research operates within specific boundaries that define its focus while acknowledging the broader context of public health in Thailand:

1.5.1 Geographical Context: The study is confined to Thailand's provincial administrative divisions, utilizing aggregated provincial-level data rather than individual or sub-provincial metrics. This approach aligns with Thailand's healthcare planning structures while providing sufficient granularity to identify regional patterns in disease distribution and social determinants.

1.5.2 Temporal Boundaries: The analysis spans a decade (2012-2021), capturing recent trends while maintaining sufficient historical context for pattern recognition. This timeframe encompasses significant developments in Thailand's healthcare system, including the advancement of universal coverage and evolving patterns of urbanization that influence both social determinants and disease prevalence.

1.5.3 Disease Selection Criteria: The research focuses specifically on six noncommunicable conditions (DM, HTN, COPD, CVD, Stroke, and Cancer) selected based on three criteria: (a) their significant contribution to Thailand's disease burden as measured by mortality and disability; (b) their established sensitivity to social determinants as documented in international literature; and (c) the availability of reliable prevalence data across multiple years at the provincial level.

1.5.4 Data Source Parameters: The study utilizes only publicly available datasets from official government repositories, including the Ministry of Public Health (MOPH), National Statistical Office (NSO), and Pollution Control Department. This approach ensures data authority while demonstrating how existing public data resources can be leveraged for advanced health analytics without requiring costly primary data collection.

1.5.5 Methodological Boundaries: The analytical approach is limited to predictive modeling rather than causal inference, focusing on identifying patterns and associations rather than establishing definitive causal relationships. While the findings may suggest causal pathways, formal causal claims would require different methodological approaches beyond this study's scope.

1.5.6 Feature Scope Limitations: The SDH features included are constrained by data availability in public repositories, meaning some potentially relevant social determinants (such as detailed measures of social cohesion, political voice, or discrimination) are not captured. The study works within these constraints while maximizing the use of available data across all five SDH domains.

1.5.7 Application Focus: The research aims to produce insights at the population and policy level rather than for individual clinical decision-making. The predictive models are designed to inform resource allocation, intervention planning, and policy development rather than individual risk assessment or diagnosis.

CHAPTER 2

LITERATURE REVIEW

This chapter offered a thorough review of previous research about the interplay between SDH and NCD, encompassing an analysis of predictive algorithms employed in NCD forecasting and a succinct summary of unresolved issues related to feature selection within SDH that require further investigation and was published (Bhoothookngoen & Sanchan, 2023).

2.1 The Relation of SDH and NCD

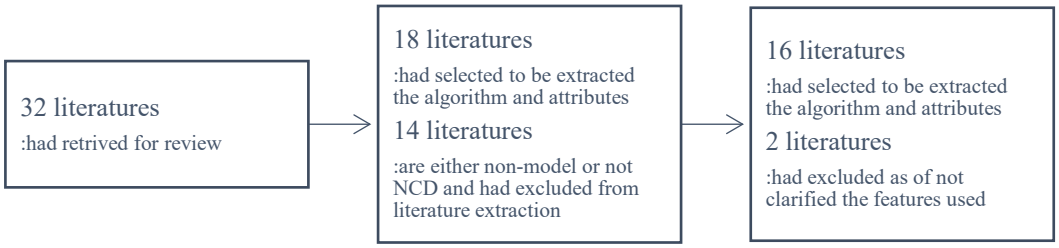
The WHO has published several documents on the interplay between SDH and NCD. The 2010 discussion paper emphasised the role of policymakers, the need for a customised conceptual framework, and the use of social and political science in methodology development. The WHO Regional Office for Europe published the "Action Plan for the Prevention and Control of Non-communicable Diseases in the WHO European Region" in 2016, aligned with SDG and UN high-level NCD meetings. Kathirvel and Thakur (2018) identified "best buy" interventions, while Potempa et al. (2022) proposed recommendations to increase Quality-Adjusted Life Year (QALY) and reduce the NCD burden in Thailand. Likewise, Urwannachotima (2016) studied SDH in Thailand, highlighting the impact of societal inequities. Nawamawat et al. (2019) found a 14.8% prevalence of NCD in a semi-urban Thai community, identifying various risk factors.

The WHO's significant attention towards the intersection of SDH and NCD has led to numerous policy frameworks, interventions, and recommendations. The studies in this section have provided valuable insights into the prevalence and risk factors of NCD in Thailand and the potential interventions to mitigate the NCD burden and increase QALY.

2.2 Existing Predictive Algorithms for NCD

The increasing prevalence of NCD has led to a growing interest in developing predictive models that accurately forecast their occurrence. A literature review was conducted to consolidate criteria-met predictive model studies for forecasting NCD prevalence to consolidate the existing literature on this topic. Out of thirty-two studies retrieved using systematic search, fourteen were excluded because of either non-model or not applied in NCD studies, and two more were excluded due to not being machine learning modelling studies, as shown in Figure 2.1. The algorithms extracted from the remaining literature included supervised, unsupervised, and reinforcement algorithms, as shown in Figure 2.2. Age was the most frequently applied attribute, followed by gender, Fasting Blood Sugar (FBS), the slope of peak exercise ST segment, Thalassemia, smoking, physical inactivity, obesity, maximum heart rate achieved, family history, exercise-induced angina, blood pressure, alcoholic, serum cholesterol, resting blood pressure, psychological stress, Diabetes Mellitus (DM), weight, ST depression induced by exercise relative to rest, and resting electrocardiographic. Evaluation methods for the models included %Accuracy, Algorithms comparison, 95% Confident Interval, Kappa stat, Root mean square error (RMSE), Precision, Recall, F measure, Receiver Operating Characteristic (ROC), and Hamming loss. Non-individual factors were used as attributes for five studies, individual factors for eleven studies, and the rest were excluded due to unidentified attributes. The studies included using individual factors proposed the outcome as an individual diagnostic result based on individual input factors. The studies by Wang Y. and Wang J. (2020), Stringhini S. et al. (2018), George N. and Thomas J. (2018), Hastings K. et al. (2022), and Hu et al. (2020) are examples of studies that used predictive models for forecasting NCD. These studies applied different algorithms, attributes, and evaluation methods to obtain the desired outcomes.

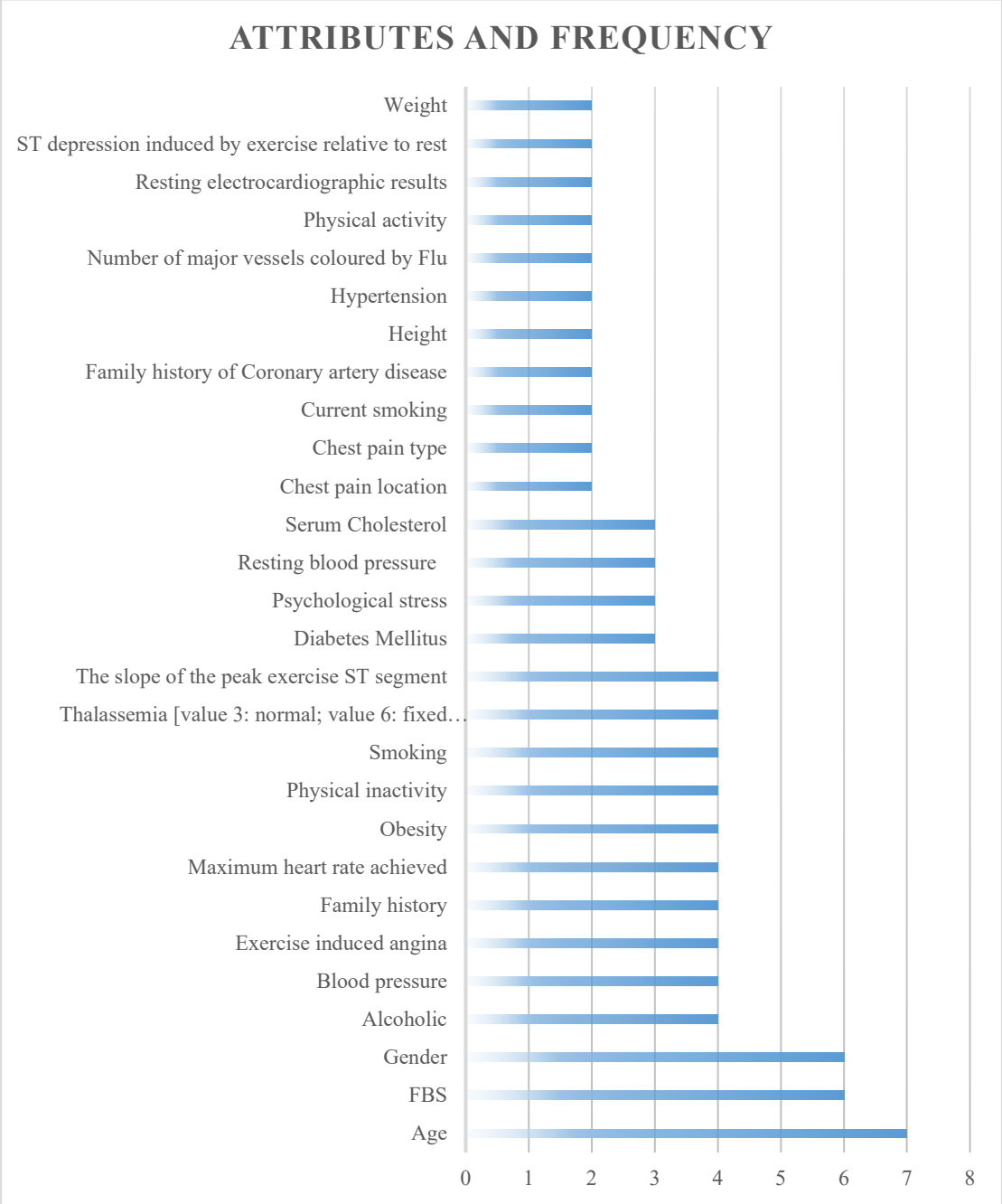
Figure 2.1: Literature Review Consort



Source: Published in Bhoothookngoen, & Sanchan, 2023.

In recent years, there has been a growing interest in utilising machine learning algorithms for predicting and diagnosing non-communicable diseases. This review summarises several studies in this area and categorises them according to the types of machine learning algorithm employed. Specifically, this thesis considers sixteen relevant studies, including the author's name, year of publication, study title, the algorithm used, model evaluation, and attributes selected for model training. This literature review finding has been published in Bhoothookngoen, & Sanchan (2023)'s article.

Figure 2.2: Feature Analysis & Deployment Rates (16 Studies)



*Showing only the attributes with frequency more than 2 times of deployment.

Table 2.1: Algorithms Used by Category

Supervised	Unsupervised	Other
Artificial Neural Network (ANN) Support Vector Machine (SVM) Long Short-Term Memory Networks (LSTM) Logistic Regression Decision Tree Naive Bayes Random Forest K-Nearest Neighbor (KNN) AdaBoost	K-means clustering Maximal Frequent Itemset Algorithm (MAFIA) Binary Relevance (BR) Classifier Chains (CC) The random k-labelsets (RAkEL) Multi-Label k-Nearest Neighbor (ML-KNN)	Deep Shapley Additive Explanations (DeepSHAP) Gradient boosting decision tree (GBDT) Combination of evolution tree model and Multilevel Modelling (MLM) Generalised Additive Mixed Model (GAMM) Fuzzy Logic IF-THEN rules Dynamic population model – Regression

Source: Published in Bhoothookngoen, & Sanchan, 2023

2.2.1 Supervised Methods

Serveralstudies have employed supervised learning techniques to predict and model NCD. For example, Ngom et al. (2020) and Saiful et al. (2020) used techniques such as Artificial Neural Networks (ANN), SVM, decision trees, Naive Bayes, logistic regression, and random forest. Additionally, Keerthi Samhitha B. et al. (2020) and Mohan N. et al. (2021) utilised decision trees, K-nearest neighbour (KNN), K-means clustering, AdaBoost, and logistic regression in their supervised learning models. In another study, Hu et al. (2018) employed a GBDT to predict non-communicable diseases and improve intervention programs in Bangladesh.

Moreover, Hu et al. (2020) utilised a stacking ensemble model that combined linear regression, support vector regression, extreme gradient boosting, random forest, and GBDT to predict daily hospital admissions for cardiovascular diseases. These studies demonstrate the effectiveness of supervised learning techniques in predicting and modelling NCD.

2.2.2 Unsupervised Methods

Banu, M. A. N., & Gomathy, B. (2014) deployed various unsupervised learning techniques, including K-means clustering, Maximal Frequent Itemset Algorithm (MAFIA), and C4.5 algorithm (supervised), to forecast the NCD. Similarly, Sangkatip and Phuboon-ob (2020) employed multiple techniques, such as binary relevance (BR), classifier chains (CC), random k-labelsets (RAkEL), and multi-label k-nearest neighbour (ML-KNN), to classify NCD. In contrast, Davagdorj et al. (2021) deployed a combination of both supervised and unsupervised learning techniques, including hybrid feature selection, eXtreme Gradient Boosting (XGBoost), logistic regression (supervised), random forest (supervised), KNN (supervised), Support Vector Machine - Recursive Feature Elimination (SVM-RFE) (supervised), Multi-Layer Perceptron (MLP) (supervised), Neural Network (NN) (supervised), and random forest-based feature selection, in their models to predict smoking-induced NCD.

2.2.3 Other Methods

Several studies have utilised various machine-learning techniques to investigate NCD. For instance, George N. and Thomas J. (2018) developed fuzzy logic-based IF-THEN rules to forecast peak demand days of chronic respiratory diseases. Hu et al. (2018) and Hu et al. (2020) also employed machine learning techniques to examine NCD. Hastings et al. (2022) utilised a dynamic population model with regression to project new-onset cardiovascular disease by socioeconomic group in Australia.

Davagdorj et al. (2021) used an Explainable Artificial Intelligence Based Framework for Non-Communicable Diseases Prediction, incorporating Deep Shapley Additive Explanations (DeepSHAP) to enhance interpretability. Wang and Wang (2020) combined the evolution tree model and Multilevel Modelling (MLM) with modeling and predict global non-communicable diseases. Lastly, Stringhini S. et al. (2018) applied a generalised additive mixed model (GMM) to study non-communicable disease risk factors in older adults. These studies collectively demonstrate the diverse machine-learning techniques employed in investigating non-communicable diseases.

Table 2.2: Forecast Model Review: Individual Attributes

Years	Authors	Algorithms	Model Evaluations	Attributes
2020	Ngom, F., Fall, I. S., Camara, M., & Bah, A. (Ngom et al., 2020)	ANN, SVM, LSTM, Decision tree, Naive Bayes, Random Forest	% Accuracy	Cardiovascular comorbidity, Drug used, Chest pain type, Height, Hypertension, Number of major vessels coloured by Flu, Resting electrocardiographic results, ST depression induced by exercise relative to rest, Weight, Diabetes Mellitus, Exercise-induced angina, Maximum heart rate achieved, Psychological stress, Resting blood pressure, Thalassemia, The slope of the peak exercise ST segment, Alcoholic, Family history, Obesity, Physical inactivity, Smoking, FBS, Gender, Age
2021	Ferdousi, R., Hossain, M. A., & El Saddik, A. (Ferdousi et al., 2021)	Random Tree	Kappa statistic, Root Mean Square Error (RMSE), TP Rate, FP Rate, Precision, Recall, F-measure, Receiver Operating Characteristic (ROC), Accuracy	Alopecia, Delayed healing, Genital thrush, Irritability, Itching, Muscle stiffness, Partial paresis, Polydipsia, Polyphagia, Polyuria, Sudden weight loss, Visual blurring, Weakness, Obesity, Gender, Age
2020	Islam, S., Jahan, N., & Khatun, M. E. (Islam et al., 2020)	Logistic regression, Decision tree, SVM, Naive Bayes	Compared with the UCI dataset result, % Accuracy	Blood cholesterol, Diet, Physical Activity, Blood pressure, Diabetes Mellitus, Psychological stress, Age, Obesity, Alcoholic, Family history

(Continued)

Table 2.2 (Continued): Forecast Model Review: Individual Attributes

Years	Authors	Algorithms	Model Evaluations	Attributes
2014	Banu, M. A. N., & Gomathy, B. (Banu& Gomathy et al, 2014)	K-means clustering, MAFIA, C4.5 Algorithm	Precision, Recall,, Accuracy	Patient Id, Age, Gender, The slope of the peak exercise ST segment, family history of coronary artery disease, Fasting Blood Sugar, chest pain location, Thalassemia, serum cholesterol, resting blood pressure, exercise induced angina, Maximum Heart Rate Achieved
2020	Alim, M. A., Habib, S., Farooq, Y., & Rafay, A. (Alim et al., 2020)	Random forest, Stratified Kfold	% Accuracy compared with other algorithm e.g., Logistic regression, SVM, Naive Based, Gradient Boosting	Target class, Chest pain type, Number of major vessels coloured by Flu, resting electrocardiographic results, Serum Cholesterol, ST depression induced by exercise relative to rest, Exercise induced angina, Maximum heart rate achieved, Resting blood pressure, Thalassemia, The slope of the peak exercise ST segment, FBS, Gender, Age
2020	Worawith Sangkatip, Jiratta Phuboon-ob (Songkatip& Phuboon-ob, 2020)	Binary Relevance (BR), Classifier Chains (CC), The random k-labelsets (RAkEL), Multi-Label k-Nearest Neighbor (ML-KNN)	% Accuracy, Hamming Loss	Diagnosis, Height, Weight, Blood pressure, Alcoholic, Family history, Smoking, FBS

(Continued)

Table 2.2 (Continued): Forecast Model Review: Individual Attributes

Years	Authors	Algorithms	Model Evaluations	Attributes
2020	Keerthi Samhitha, B., Sarika Priya., M., Sanjana., C., Mana, S. C., & Jose, J. (Samhitha et al., 2020)	Decision tree, KNN, K-means clustering, AdaBoost	Exactness, Accuracy, Mistake in grouping	The 13 characteristics of the informational collection
2021	Davagdorj, K., Bae, J. W., Pham, V. H., Theera-Umpon, N., & Ryu, K. H. (Davagdori et al., 2021)	Deep Shapley Additive Explanations (DeepSHAP)	Accuracy, Specificity Recall (Sensitivity), Precision, F Scores, AUC	BMI, Education level, how often add salt to food at table, Insomnia, Marital Status, Monthly poverty level of family, Past year Doctor visit frequency, Poor appetite or overeating, Pulse regularity, Taking insulin now, Total number of people in household, Current smoking, Physical activity, Blood pressure, psychological stress, Alcoholic, Family history, Physical inactivity, FBS, Gender, Age
2021	Mohan, N., Jain, V., & Agrawal, G. (Mohan, Jain, and Agrawal, 2021)	KNN, Naive Bayes, Random Forest, Logistic Regression	None	Chest pain location, Family history of Coronary artery disease, Exercise induced angina, Maximum heart rate achieved, Resting blood pressure, Thalassemia, The slope of the peak exercise ST segment, FBS, Gender, Age

(Continued)

Table 2.2 (Continued): Forecast Model Review: Individual Attributes

Years	Authors	Algorithms	Model Evaluations	Attributes
2020	Davagdorj, K., Pham, V. H., Theera-Umpon, N., & Ryu, K. H. (Davagdori et al., 2021)	XGBoost, Hybrid Feature Selection (HFS), Logistic regression, Random Forest, KNN, SVM-RFE, MLP, NN, Random Forest based feature selection - RFFS	Compared with baseline model	Gender, Age, Household income, Education, Occupation, Marital status, Subjective health status, Depression diagnosis, Health checkup status, Athletic ability, Self-management, Daily activities, Pain/discomfort, Anxious/Depressed, EQ-5D index, Economic activity status, Weight control: exercise, Lifetime drinking experience, Start drinking age, Frequency of drinking for 1-year, Monthly drinking rate, Stress level, Indoor indirect smoking exposure, The usual time spent sitting (day), Walk duration (hours), Family history of chronic disease, BMI, Obesity prevalence, FBS, Total cholesterol, Flexible exercise days per week, Residence area, etc.
2018	Hu, M., Nohara, Y., Wakata, Y., Ahmed, A., Nakashima, N., & Nakamura, M. (Hu et al., 2018)	Gradient boosting decision tree (GBDT)	AUC, Sensitivity, Specificity, f measure, Accuracy	literacy, occupation, time since the last meal, present symptoms, past diseases, medication, smoking, weight change, exercise, walking speed, eating behavior, sleeping, and the desire to have a healthy lifestyle, subjects underwent a health checkup using the sensor devices, blood glucose, blood pressure, weight, height, etc.

Source: Published in Bhoothookngoen, & Sanchan, 2023

Table 2.3 Forecast Models: Review of Non-Individual Attributes

Years	Authors	Algorithms	Model Evaluations	Attributes
2020	Wang, Y., & Wang, J. (Wang Y & Wang, 2020)	Combination of evolution tree model and MLM	Accuracy compared with Linear regression	Country type [income] and Country development stage, NCD death, Socio-economic status
2018	Stringhini, S., Carmeli, C., Jokela, M., Avendaño, M., McCrory, C., d'Errico, A., . . . Kivimäki, M. (Stringhini et al., 2018)	Generalised additive mixed model (GAMM)	5000 bootstrap samples, 95% CI	High alcohol intake, Low socioeconomic status, Current smoking, Hypertension, Diabetes Mellitus, Obesity, Physical inactivity
2018	George, N., & Thomas, J. (George& Thomas, 2018)	IF-THEN rules	Compared with the original data	Nitrogen Dioxide, Outdoor temp, Particle matter, Relative humidity, Sulphur Dioxide, Wind speed

(Continued)

Table 2.3 (Continued): Forecast Models: Review of Non-Individual Attributes

Years	Authors	Algorithms	Model Evaluations	Attributes
2020	Hu, Z., Qiu, H., Su, Z., Shen, M., & Chen, Z. (Hu et al., 2020)	Linear regression, Support vector regression, Extreme gradient boosting, Random forest, Gradient boosting decision tree	Mean absolute error (MAE), Root mean square error (RMSE), Mean absolute percentage error (MAPE), Coefficient of determination (R square)	Air quality, Hospital admission, Meteorological
2022	Hastings, K., Marquina, C., Morton, J., Abushanab, D., Berkovic, D., Talic, S., Zomer, E., Liew, D., & Ademi, Z. (Hastings et al., 2022)	Dynamic population model - Regression	Sensitivity analysis	Population, Risk of new-onset CVS by socioeconomic quintile, Utility

Source: Published in Bhoothookngoen, & Sanchan, 2023

2.3 SDH-Related Features in Previous Works

The experiment aimed to utilize SDH datasets as features for the prediction model of NCD prevalence. In this section, the results from previous studies exploring SDH will guide the approach. From various recent studies have been conducted to develop predictive models for NCD using different attributes. In eleven studies, individual factors were chosen as the attributes for model training, while five chose non-individual factors. The remaining two studies did not identify the selected features in the retrieved full-text articles. The five studies that selected non-individual factors as features for model training, as summarized in Table 2.4, demonstrated varying coverage across the five SDH domains.

Wang Y. and Wang (2020) studied the predictive model for global NCD deaths deploying the socioeconomic factors, country development level, income (country level), and the number of NCD deaths. They proposed a novel algorithm combining the evolution tree and Multilevel models (MLM). They compared the proposed algorithm with LR and found that the R square value was 0.7932 for the proposed novel model and 0.7005 for LR. The study found an association between socioeconomic factors and NCD death.

Stringhini et al. (2018) studied the association between low socioeconomic status and NCD risk factors such as diabetes, high alcohol intake, high blood pressure, obesity, physical inactivity, and smoking among older individuals in multi-cohort populations from 24 countries. The authors used generalised additive mixed models (GAMM) for analysis and found an association between socioeconomic status and physical functioning.

George and Thomas (2018) studied a model for forecasting the peak demand days of chronic respiratory diseases using Fuzzy logic. They applied environmental factors to predict the peak demand day and evaluated the model by comparing it with the original data.

Hastings et al. (2022) deployed a dynamic population model to determine Australia's new-onset cardiovascular disease (CVD) by socioeconomic group. The simulation included population, risk of new-onset CVD by socioeconomic quintile,

and utility. The study found that 8.4% of people in the most disadvantaged quintile were at high risk of CVD.

Hu et al. (2020) studied a predictive model for the number of CVD admissions using air quality, hospital admission, and meteorological data. The stacking model and Sequential Forward Floating Selection (SFFS) for feature selection were deployed in model training. The results have been evaluated by comparing the MAE, RMSE, MAPE, and R square of RF (the second stage) with the MAE, RMSE, and MAPE of the first stage algorithms, decreasing by 6.3%, 7.4%, and 6.3%, respectively, and the R2 improving by 1.7%, compared with the performance of the second stage (RF) for the final prediction result.

The reviewed studies have used various attributes to develop predictive models for NCD, with individual factors being the most commonly used. As shown in Table 5, while these studies incorporated some SDH domains, particularly economic stability and neighbourhood environment, they notably lacked comprehensive coverage across all five SDH domains. The education access and quality domain were entirely absent, while healthcare access and quality and social and community context were minimally represented. However, insights from these five studies using non-individual factors highlight the association between socioeconomic and environmental factors and NCD. Therefore, it is important to consider different attributes when developing NCD predictive models. Based on the similarity in study design and selected attributes, Hu et al.'s (2020) study is particularly relevant to SDH most, and its model development method will be modified to use for this study.

Table 2.4: Previous Studies' Attribute Coverage Across SDH Domains

Study	Social Determinants of Health				
	Economic Stability	Education Access & Quality	Healthcare Access & Quality	Neighbourhood & Built Environment	Social & Community Context
Wang & Wang (2020)	Country income level (worldwide)	-	-	-	Country development stage
Stringhini S et al. (2018)	Socioeconomic status	-	-	-	
George & Thomas (2018)	-	-	-	Air quality, Outdoor temperature, Relative humidity, Wind speed	-
Hastings K et al. (2022)	Socioeconomic quintile	-	-	-	-
Hu et al. (2020)	-	-	Hospital admission data	Air quality, Meteorological data	-

As demonstrated in Table 2.4, the five previous studies explored only select aspects of socioeconomic and environmental factors in NCD prediction, with none comprehensively covering all SDH domains. The significant gap in literature lied in the absence of a predictive model that integrated all five domains of SDH - economic stability, education access and quality, healthcare access and quality, neighbourhood and built environment, and social and community context. This thesis aimed to address this gap by leveraging insights from prior research, particularly the methodology employed by Hu et al. (2020). While Hu et al.'s study focused primarily on healthcare access and environmental factors, this study adapted and modified their model development approach to incorporate all five SDH domains as features. This modification aimed to enhance the accuracy and comprehensiveness of the predictive model for NCD prevalence, providing a more holistic approach to understanding the social determinants influencing NCD in Thailand.

2.4 Rationale for Adopting the Stacking Ensemble Methodology

The selection of a stacking ensemble methodology for this study was strategically aligned with the research questions and objectives based on several critical considerations. Stacking, as implemented by Hu et al. (2020), offered distinct advantages for investigating the complex relationships between SDH and NCD prevalence in Thailand.

First, stacking ensemble models excel at handling diverse predictor variables across different scales and distributions crucial capability when working with heterogeneous SDH features spanning economic, educational, healthcare, environmental, and social domains. Each base learner in the ensemble captured different aspects of the relationship between these features and disease outcomes, mitigating the limitations any single algorithm might encounter with Thailand's complex socioeconomic landscape.

Second, the research questions explicitly sought to evaluate which specific SDH features demonstrated the highest predictive importance across different NCD categories. Ensemble methods typically provide more robust feature importance

assessments than single models, as they reduce the impact of algorithm-specific biases in feature evaluation. By implementing multiple base learners (LR, SVR, XGBoost, RF, and GBDT) with RF as a meta-learner, this study could triangulate feature importance across different algorithmic perspectives, addressing the second research question with greater validity.

Third, previous research had not systematically compared different data preprocessing strategies when working with spatially inconsistent SDH datasets, as specified in the third research question. The stacking approach, with its inherent cross-validation methodology, provided an ideal framework for systematically evaluating how different preprocessing strategies affected model performance across multiple disease categories and heterogeneous data distributions.

Fourth, Hu et al.'s (2020) implementation demonstrated successful application in a similar context, predicting daily cardiovascular hospital admissions using environmental and healthcare utilization data. While their study incorporated only two SDH domains (healthcare access and neighbourhood environment), their methodological approach offered a validated foundation that could be extended to incorporate the comprehensive five-domain SDH framework examined in this study.

Finally, stacking ensemble methods typically demonstrate superior performance when working with datasets characterized by complex, non-linear relationships and inconsistent data quality precisely the challenges presented by Thailand's provincial-level SDH and NCD prevalence data. The method's ability to handle missing values and spatial inconsistencies through its hierarchical learning structure made it particularly suitable for addressing the first research question regarding the overall effectiveness of SDH features in predicting NCD prevalence.

These considerations collectively justified the adaptation of Hu et al.'s (2020) stacking ensemble approach as the methodological foundation for this study, enabling a comprehensive investigation of how SDH features predict NCD prevalence patterns across Thailand's provinces.

CHAPTER 3

METHODOLOGY

In the development of machine learning models for healthcare predictions, particularly for NCD, the quality and preparation of data form the cornerstone of reliable outcomes. This chapter systematically explored the data collection, processing, and preparation methodologies employed to ensure robust analysis and meaningful results. Through careful consideration of data sources, cleaning procedures, and feature engineering, this study established a solid foundation for the predictive modelling of NCD in Thailand using SDH features.

3.1 Software Environment

The methodology started with the development environment setup, where Visual Studio Code (v1.75) serves as the primary integrated development environment. Python was selected as the core programming language for model development and statistical analysis, complemented by Jupyter Notebook for interactive code development and result visualization. Microsoft Excel (v16.69.1) facilitated initial data cleaning, transformation, and exploratory analysis, while essential Python libraries including scikit-learn, pandas, and numpy were configured for machine learning operations.

3.2 Data Collection Framework

3.2.1 Strategic Approach

The investigation of relationships between NCD and their social determinants required a comprehensive and methodologically sound data collection strategy. This is particularly crucial in Thailand, where the interplay between SDH and disease outcomes presented unique challenges for public health research. Given these considerations, this thesis employed a systematic approach to data collection, encompassing multiple dimensions of both health outcomes and their social determinants.

The data collection strategy was designed to address three key objectives including capturing comprehensive longitudinal data on NCD prevalence, gathering detailed information on various SDH factors across different regions, and ensuring data quality and reliability through the use of authoritative sources.

To achieve these objectives, primary data collection was conducted between September and December 2022, utilizing a multi-source approach. This timeframe was strategically chosen to capture the most recent available data while ensuring sufficient historical context for trend analysis. The selection of data sources prioritized official government repositories and validated databases to maintain data integrity and reliability.

3.2.2 Data Source Overview

This thesis utilized various data sources to investigate the relationships between SDH, NCD, and hospital outcomes in Thailand. The data were collected over multiple years to comprehensively understand the patterns and trends. Primary data collection occurred between September and December 2022, drawing from several authoritative sources as outlined in Table 3.1.

The Open Data web portal by The Ministry of Public Health (MOPH) provided a comprehensive dataset on various health-related aspects of Thai citizens. This included detailed information on health service access, healthcare providers, health status by various diseases (including NCD), cause of illness, Tuberculosis related activities, and diseases from occupation or environment.

The MOPH Open Data web portal served as the primary source of information for the number of NCD patients for at least a decade before 2021.

Table 3.1: Data Sources and Strategic Approach

Datasets	Date of Retrieval	Source	Period Covering	Column Features
Prevalence of Breast Cancer (BA)φ	10 September 2022	Open Data by MOPH	2013 - 2021	Hospital code and different age groups
Prevalence of Cardiovascular disease (CVD) φ				
Prevalence of Cervical cancer (CC) φ				
Prevalence of Chronic Obstructive Pulmonary Disease (COPD) φ				
Prevalence of Diabetes Mellitus (DM) φ				
Prevalence of Hypertension (HTN) φ				
Prevalence of Lung cancer (LC) φ				
Prevalence of Stroke (Cerebrovascular disease) φ				

(Continued)

Table 3.1 (Continued): Data Sources and Strategic Approach

Datasets	Date of Retrieval	Source	Period Covering	Column Features
Amount of Household Income[ε]		National Statistical Office (NSO)	2004, 2006, 2007, 2009, 2011, 2013, 2015, 2017, 2019, 2021	Province
Amount of Household Expense[ε]			2012 – 2021	Province
Number of years earned education[ε]				Gender and different age groups
Number of educational institutions[ε]			2016 – 2021	Divided by Bangkok vs Upcountry group
Number of Smoker[ε]			2013-2014, 2017, 2021	Gender, different age groups, and region
Number of Alcoholic Consumer[ε]			2009, 2011, 2013 – 2015, 2017, 2021	Gender
The number of PM 2.5[ε]		Air 4 Thai Website	2013 - 2021	Province
Hospital List[ε]	12 November 2022	MOPH's website	2016 and backwards	5 digits code, 8 digits code, hospital type, hospital name, higher government body code, area code, address, operating status

ε - SDH datasets, φ - NCD prevalence datasets

3.2.2.1 NCD prevalence data collection

The data were retrieved between September and December 2022 from internet access via various primary sources: the Open Data web portal by MOPH (Open Government Data of Thailand, n.d.), the Pollution Control Department (Pollution Control Department, 2022), NSO's website (National Statistical Office Thailand, n.d.) and MOPH's website (The Ministry of Public Health of Thailand, 2016).

The Open Data web portal by MOPH provided a comprehensive dataset on various health-related aspects of Thai citizens. This included information on health service access, healthcare providers, health status by some diseases (including NCD), cause of illness, Tuberculosis related activities, and diseases from occupation or environment. The MOPH Open Data web portal was the primary source of information for the number of NCD patients for at least a decade before 2021.

Prevalence rates of specific NCD were obtained from 2012 to 2021, including Breast Cancer (BC), Cardiovascular Disease (CVD), Cervical Cancer (CC), Chronic Obstructive Pulmonary Disease (COPD), Diabetes Mellitus (DM), Hypertension (HTN), Lung Cancer (LC), and Stroke. Finally, hospital lists matched the number of patients to each hospital code and determined the locations of the particular hospitals in different provinces.

3.2.2.1 SDH data collection

The SDH were categorized according to the Healthy People 2030 framework (Healthy People 2030 | health.gov, n.d.), encompassing five key domains including economic stability, education access and quality, health care access and quality, neighbourhood and built environment, and social and community context.

The NSO is a government body under the Ministry of Digital Economy and Society of Thailand that manages statistical information to support economic development and competitiveness. The NSO website was the source of data on SDH, including household income, expenses, loans, years of education, number of educational institutions, number of smokers, and number of alcoholic consumers.

At the same time, the levels of particulate matter 2.5 could be found on the Pollution Control Department's website (Pollution Control Department, 2022), Ministry of Natural Resources and Environment of Thailand. Data on household

income, expenses, and loans were collected to assess Economic Stability. Education Access and Quality were evaluated by collecting data on the years of education and the number of educational institutions in each province.

Data was collected on the number of hospitals for each province to evaluate Health Care Access and Quality. The Neighbourhood and Built Environment domain was assessed by collecting data on the levels of particulate matter 2.5 (pm2.5) in each province. Lastly, Social and Community Context was assessed using hospital location (province) data. By assessing these SDH domains, this study aimed to identify potential associations between SDH factors and the prevalence of NCD in Thailand.

3.3 Dataset Characteristics

Following the data cleaning process, the individual datasets were merged into a single dataset that is now fully prepared for the subsequent stages of this study: modelling development. The merged dataset comprises 24 columns, as listed in Table 3.2; The dataset has 1,971,897 rows and 26 columns, containing 47,325,528 cells. Out of these cells, 5,630,138 are "0" (Zero). This detailed and extensive dataset will be the foundation for analysing the research problem under investigation.

The data characteristics across different NCD datasets revealed varying patterns of completeness and scale, as shown in Table 7. Cardiovascular Disease (CVD) dataset contained 532,297 rows with 5,855,278 cells, where 20% (1,213,606) of values were missing. The COPD dataset comprised 464,923 rows with 5,114,164 cells, showing 22% (1,123,534) missing values. For Stroke, the dataset included 616,151 rows and 6,777,672 cells, with 23% (1,541,882) missing values. The Diabetes Mellitus (DM) dataset was larger, containing 722,377 rows and 7,946,158 cells, though it had a higher proportion of missing values at 32% (2,565,904). Similarly, the Hypertension (HTN) dataset was substantial with 729,756 rows and 8,027,327 cells, also showing 32% (2,591,669) missing values. The Cancer (CA) dataset was comparatively smaller, with 244,868 rows and 2,693,559 cells, where 21% (556,091) of values were missing. These patterns of missing data and dataset sizes significantly influenced the subsequent approaches to data preprocessing and model development.

Table 3.2: Overview of Data Characteristics

General Description				
CVD				
Number of rows	532,297		Number of missing values (%)	1,213,606 (20%)
Number of cells	5,855,278			
COPD				
Number of rows	464,923		Number of missing values (%)	1,123,534 (22%)
Number of cells	5,114,164			
Stroke				
Number of rows	616,151		Number of missing values (%)	1,541,882 (23%)
Number of cells	6,777,672			
DM				
Number of rows	722,377		Number of missing values (%)	2,565,904 (32%)
Number of cells	7,946,158			
HTN				
Number of rows	729,756		Number of missing values (%)	2,591,669 (32%)
Number of cells	8,027,327			
CA				
Number of rows	244,868		Number of missing values (%)	556,091 (21%)
Number of cells	2,693,559			

3.3.1 NCD Prevalence Dataset Overview

In Table 8, which shows the top 10 highest counts of NCD patients from 2012-2021 in Thailand, a clear pattern emerges in the prevalence of Hypertension across different provinces and years. The data reveals that Hypertension was consistently the most prevalent condition, with the highest recorded number being 9,558 patients in 2017 (location missing), followed by 9,170 patients in Ubonratchathani province in 2014. Other significant records included 5,931 Hypertension cases in 2016 (location missing), 5,857 cases in Khonkaen (2012), and 5,848 cases in Khonkaen (2013). The provinces of Phangnga, Chonburi, and Ubonratchathani also reported notable numbers of Hypertension cases, with patient counts ranging from 4,919 to 5,723, demonstrating the widespread and significant burden of this condition across different regions of Thailand during this period.

Table 3.3 Top 10 Highest Patients (2012 – 2021)

Rank	Noncommunicable Disease	Year	Province	Number of Hospital	Number of Patients
1	Hypertension	2017	Missing	Missing	9,558
2	Hypertension	2014	Ubonratchathani	422	9,170
3	Hypertension	2016	Missing	Missing	5,931
4	Hypertension	2012	Khonkaen	373	5,857
5	Hypertension	2013	Khonkaen	373	5,848
6	Hypertension	2014	Phangnga	89	5,723
7	Hypertension	2013	Chonburi	319	5,199
8	Hypertension	2017	Chonburi	319	5,104
9	Hypertension	2017	Missing	Missing	4,924
10	Hypertension	2014	Ubonratchathani	422	4,919

3.3.2 SDH Dataset Overview

Regarding the SDH dataset, the result was published as per Bhoothookngoen & Sanchan (2024) demonstrated in Table 3.4, presenting various SDH influencing the health status of individuals in Thailand from 2013 to 2021. The data covers household income, expenses, loans, education, health, and environmental factors. Key columns include "Household Income (THB/month)," "Household Expense (THB/month)," "Household Loan (Year)," "Year of Schooling (Year)," "Educational Facility (Institution)," "Alcohol Intake (/100,000)," "Smoking Rate (/100,000)," and "pm2.5 levels (μg/m³)."

The average household income rose from 23,182 THB/month in 2013 to 24,666 THB/month in 2021, while the average loan amount increased from 148,971 years in 2013 to 202,947 years in 2021, suggesting a potential financial strain on households. Furthermore, the average concentration of pm2.5 particles in the air surged from 235 μg/m³ in 2013 to 1,677 μg/m³ in 2021, indicating a significant environmental hazard that could affect public health.

Table 3.4: Dataset Feature: SDH Feature

Year	Household Income (THB/month) (Min - Max)	Household Expense (THB/month) (Min - Max)	Household Loan (Year) (Min - Max)	Year of Schooling (Year) (Min - Max)
2013	23,182 (8,821 - 49,191)	17,731 (7,405 - 35,024)	148,971 (9,857 - 386,957)	27 (4.2 - 8.9)
2014	Not available	18,665 (9,686 - 34,426)	Not available	28 (4.3 - 9)
2015	23,542 (13,497 - 45,572)	18,982 (11,864 - 33,086)	156,346 (8,090 - 373,325)	29 (4.49 - 9.38)
2016	Not available	18 777 (11 859, 35 101)	Not available	29 (4.53 - 9.47)
2017	23,840 (11,809 - 45,707)	18,959 (10,441- 35,351)	173,535 (28,438 - 294,901)	29 (4.55 - 9.6)
2018	Not available	18,764 (11,213 - 43,301)	Not available	29 (4.6 - 9.68)

(Continued)

Table 3.4(Continued): Dataset Feature: SDH Feature

Year	Household Income (THB/month) (Min - Max)	Household Expense (THB/month) (Min - Max)	Household Loan (Year) (Min - Max)	Year of Schooling (Year) (Min - Max)
2019	23,568 (13,971 - 46,978)	18,521 (11,243 - 37,086)	157,704 (16,895 - 288,110)	30 (4.74 - 9.78)
2020	Not available	19,173 (11,532 - 33,824)	Not available	31 (4.94 - 10.02)
2021	24,666 (15,496 - 41,129)	19,500 (12,214 - 33,996)	202,947 (47,603 - 370,531)	31 (5.08 - 10.13)

Year	Educational Facility (Institution)	Alcohol Intake Rate (/100,000)	Smoking Rate (/100,000)	pm2.5 levels (µg/m³) (Min - Max)
2013	Not available	32,892	Not available	33.51 (19.46 - 62.07)
2014	Not available	32,950	Not available	28.78 (19.68 - 39.30)
2015	Not available	34,786	Not available	27.91 (16.39 - 46.23)
2016	77,258	Not available	Not available	26.93 (11.78 - 43.27)
2017	76,516	29,050	Not available	22.40 (8.89 - 35.83)
2018	76,712	Not available	Not available	23.72 (8.76 - 41.39)
2019	75,962	Not available	Not available	25.49 (9.69 - 40.57)
2020	75,475	Not available	Not available	23.27 (7.69 - 42.39)
2021	Not available	28,600	570	21.50 (9.72 - 39.59)

3.3.3 Trends Analysis

The result was published as per Bhoothookngoen & Sanchan (2024) demonstrated in Figure 3.1, 3.2, 3.3, and 3.4, regarding the prevalence of NCD in Thailand is highlighted in Figure 3.1. The data indicated a rise in patients diagnosed

with diseases such as DM, HTN, Stroke, CVD, CAs, and COPD over the examined period. The observed increases, averaging between 6.01% and 6.33% per annum, signify a growing burden of NCD on the healthcare system and population health.

In Figure 3.2, cumulative disease rates across multiple provinces in Thailand shed light on regional disparities in disease prevalence. Certain provinces exhibit significantly higher rates of DM, COPD, CAs, stroke, HTN, and CAD, underscoring the need for targeted interventions and resource allocation to address varying healthcare needs across regions.

Moreover, socio-economic and environmental factors, detailed in Figure 3.3, offer insights into the determinants of health outcomes in different provinces. Disparities in healthcare infrastructure, household finances, and environmental conditions are evident, with implications for health equity and access to healthcare services.

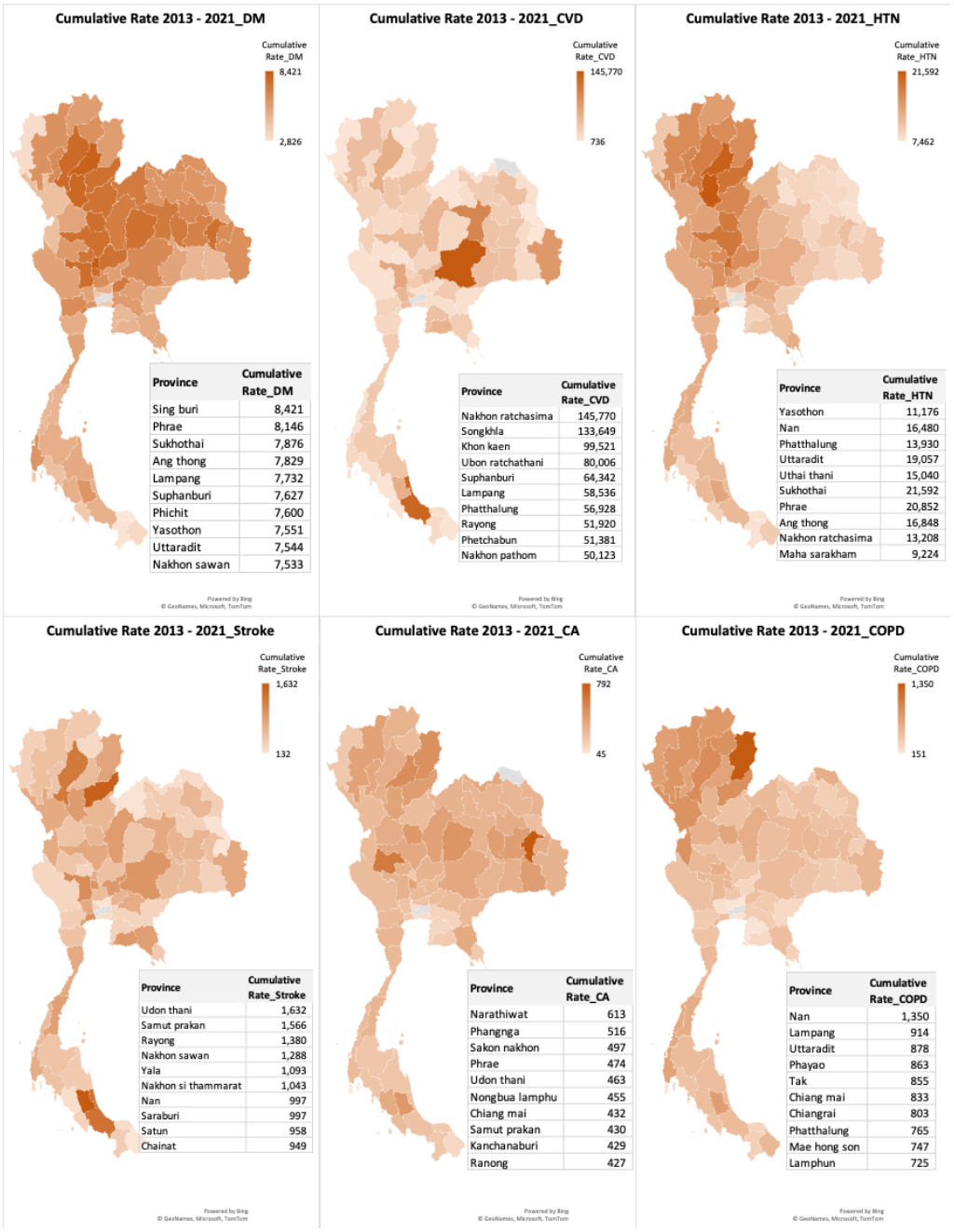
Financial trends depicted in Figure 3.4 highlight changing patterns in borrowing, expenses, and income among Thai individuals from 2004 to 2021. The observed increase in borrowing and expenses, coupled with rising income levels, reflects evolving economic dynamics and shifts in consumer behaviour over time.

Collectively, the data underscores the intricate interplay between health, socio-economic factors, and environmental conditions in shaping public health outcomes in Thailand. Addressing these multifaceted challenges requires comprehensive policy responses and targeted interventions to promote health equity, mitigate disease burden, and foster sustainable development across the country.

Figure 3.1: Analyzing the Evolution of NCD Patient Trends in Thailand

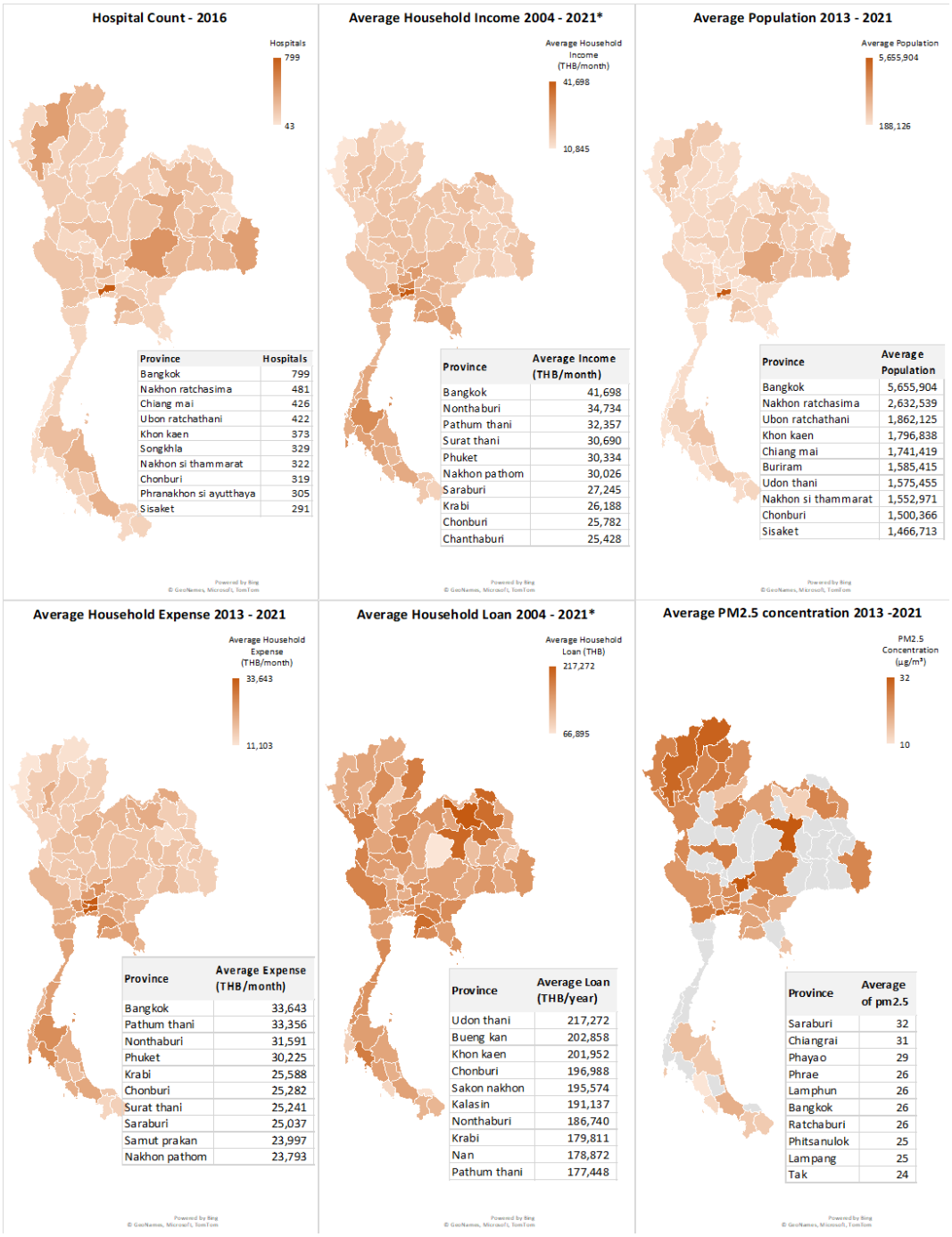
Source: Bhoothookngoen & Sanchan, 2024

Figure 3.2: Spatial Analysis of Cumulative NCD Rates in Thailand



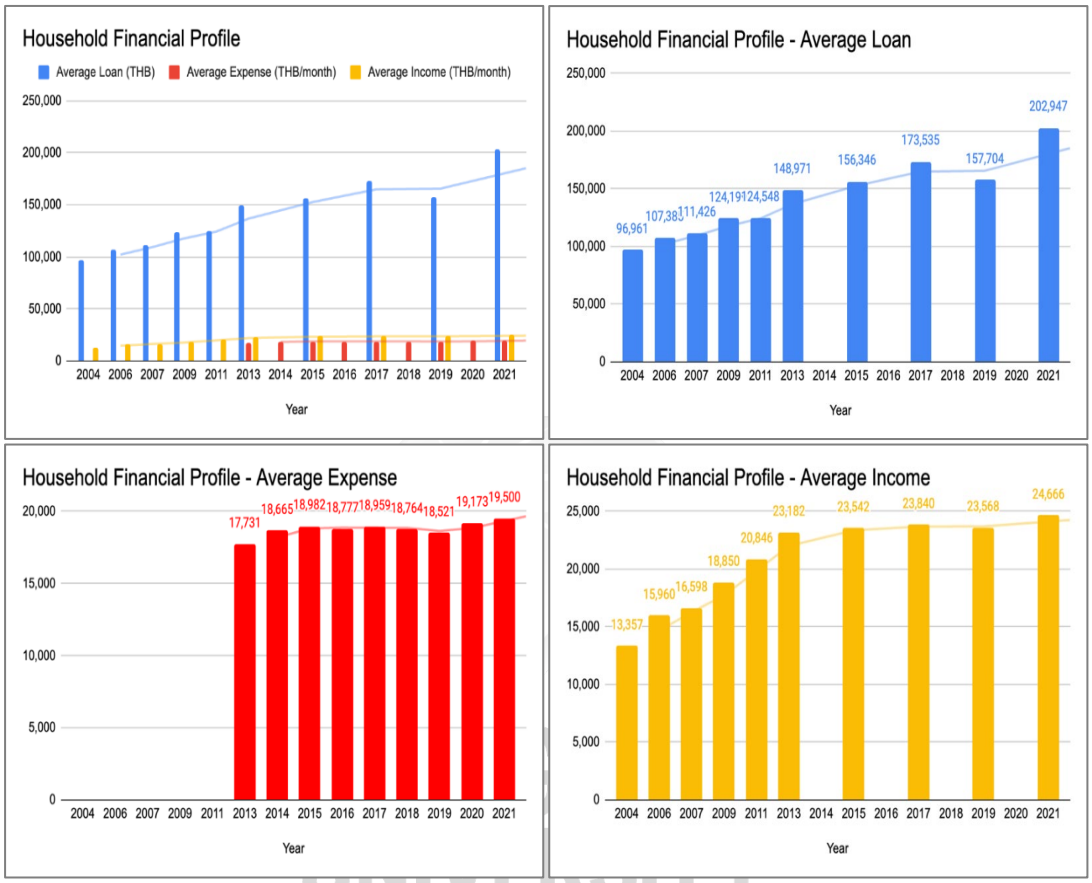
Source: Bhoothookngoen & Sanchan. (2024).

Figure 3.3: Geographic Patterns of SDH in Thailand.



Source: Bhoothookngoen & Sanchan. (2024). *Geographic patterns of social determinants of health in Thailand*. (Data available for 2004, 2005, 2007, 2009, 2011, 2013, 2015, 2017, 2019, and 2021).

Figure 3.4: Historical Trends of Household Financial Profiles in Thailand.



Source: Bhoothookngoen & Sanchan. (2024).

3.4 Data Pre-Processing

3.4.1 Data Cleaning and Manipulation

In machine learning, data pre-processing is a crucial step that directly affects the accuracy and effectiveness of trained models. Raw data may contain inconsistencies, errors, missing values, and noise that can impede machine-learning algorithms. Data cleaning eliminates these issues, resulting in more reliable and useful data. It handles missing data via imputation, reduces noise through smoothing and filtering, and transforms features via normalisation and scaling. The pre-processing of data was performed using Microsoft Excel for Mac version 16.69.1. The data cleaning process involved utilising advanced data manipulation functions, such as vlookup,

pivot tables, and other relevant techniques, to effectively remove inconsistencies and redundancies in the data.

3.4.1.1 HTN

Hypertension is a condition in which the blood pressure is higher than normal (upon defining the normal range by each hospital). In this dataset of Thai HTN patients, the data characteristics are including 729,757 rows (excluded header), 2,852,726 missing values, and 1,053,088 data points which are “0” (Zero).

3.4.1.2 CA

CA in this study is a group of diseases including Lung cancer (LC), Cervical cancer (CC), and Breast cancer (BC). In this dataset of Thai CA patients, the data characteristics are including 464,924 rows (excluded header), 1,737,501 missing values, and 1,748,197 data points which are “0” (Zero).

3.4.1.3 COPD

COPD is a group of diseases causing breathing problems due to airflow limitation. In this dataset of Thai CA patients, the data characteristics are including 244,870 rows (excluded header), 960,679 missing values, and 1,382,327 data points which are “0” (Zero).

3.4.1.4 CVD

CVD is a group of diseases related to Cardiovascular issues; in this study, it includes Stroke and Myocardial Infarction (MI). In this dataset of Thai CVD patients, the data characteristics are including 532,299 rows (excluded header), 1,880,508 missing values, and 1,951,020 data points which are “0” (Zero).

3.4.1.5 Diabetes Mellitus (DM)

DM is a condition of higher sugar levels in the bloodstream. In this dataset of Thai DM patients, the data characteristics are including 532,299 rows (excluded header), 1,880,508 missing values, and 1,951,020 data points which are “0” (Zero).

In summary for the rough descriptive overview, Table 3.5 presented a statistical summary of patient counts for six major noncommunicable diseases in Thailand from 2013 to 2021. The data revealed significant disparities in prevalence patterns across different conditions, with HTN demonstrating the highest burden (maximum of 9,558 patients, average of 69.35), followed by DM (maximum of 4,924,

average of 33.64). By contrast, CA, COPD, CVD, and Stroke showed considerably lower patient counts, with averages ranging from 1.54 to 3.84. Notably, all conditions shared a minimum value of zero, indicating geographic or temporal gaps in reporting. The consistency of the mode value (1) for four of the six conditions suggested that while occasional high-prevalence clusters occurred, most reporting units typically encountered modest case numbers for most NCDs, with HTN and DM representing the exceptions to this pattern with modes of 35 and 20, respectively.

Table 3.5: Minimum, Maximum, Average, Medium, and Mode of NCD Patients.

Noncommunicable Disease	Patient Count (2013 – 2021)				
	Minimum	Maximum	Average	Medium	Mode
CVD	0	752	3.8431	2	1
COPD	0	398	3.1805	2	1
Stroke	0	752	3.6197	2	1
DM	0	4,924	33.6366	27	20
HTN	0	9,558	69.3508	54	35
CA	0	235	1.5354	1	1

3.4.2 Missing Value

The missing value is the information missing in quantitative research, including types of missing data, the potential impact on results, and strategies for preventing and handling missing values (acceptance, deletion, and imputation) (Bhandari, 2022b).

3.4.3 Outlier

The outliers are extreme values that differ from most data points in a dataset and can result from natural variation, incorrect data entry, equipment malfunctions, or other measurement errors. There are four ways to identify outliers: the sorting method, data visualisation method, statistical test scores (z scores), and the interquartile range method. The interquartile range (IQR) method involves identifying

the first quartile, the median, and the third quartile, calculating the IQR, and using the upper and lower fences to highlight any outliers. (Bhandari, 2022a).

The detection of outliers is essential during data pre-processing; otherwise, these outliers might interfere with the accuracy of the predictive model. There are various options for detecting the outliers, such as the Box plot, Scatter plot, and mathematical functions.

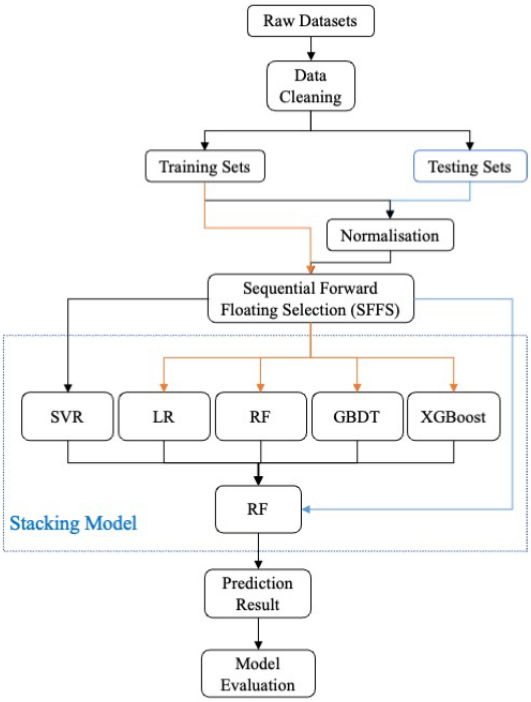
This chapter has provided a detailed account of the preliminary procedures essential to preparing datasets for modelling development in subsequent chapters. Drawing on the information and insights presented in previous chapters, including Chapters 1, 2, and 3 (this chapter), the modelling process will be informed by considerations such as the appropriateness of the expected outputs and relevant prior works. Chapter 4 will delve into the presumed model and the evaluation methods employed in the modelling process.

3.5 Model Adoption

Model development and implementation represented the third phase, centered around the construction of a sophisticated stacking ensemble model. The first stage incorporated multiple algorithms: LR established baseline predictive modeling, SVR handled non-linear relationships, Extreme XGBoost enhanced predictive accuracy, RF provided robust ensemble learning, and GBDT enabled iterative prediction improvement. The second stage employed RF for final prediction optimization. Model training procedures encompassed data splitting, parameter optimization, and integration of first-stage model predictions.

Figure 3.5 outlined the core components of the modelling process: data processing with feature engineering, model development with stacking ensemble, and validation and performance assessment. The diagram illustrated the sequential flow of these components, beginning with data preparation through Sequential Forward Floating Selection (SFFS) and normalization, progressing to the stacking ensemble model implementation with its first and second stage algorithms, and concluding with the comprehensive validation phase.

Figure 3.5 Diagram of Modelling



Source: Modified from Hu et al. (2020)

3.5.1 Algorithm Selection and Implementation

The stacking ensemble model incorporated multiple algorithms, following Hu et al.'s (2020) approach, with the diagram of modeling described in Figure 3.5. Each algorithm contributed distinct capabilities to the ensemble:

Linear Regression (LR)

$$y = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n + \in \quad (1)$$

where y is the predicted output, x_1, x_2, …, x_n re the input features, w_0, w_1, …, w_n are the coefficients (weights) to be learned, and $\in$ represents the error term.

Support Vector Regression (SVR)

$$f(x) = \sum_{i=1}^{n} \propto_i K(x, x_i) + b \quad (2)$$

Here, $\propto_i$ are the Lagrange multipliers, K(x, x_i) is the kernel function, and b is the bias term.

Extreme Gradient Boosting (XGBoost)

$$\hat{y} = \sum_{k=1}^{K} f_x(x_i), \quad f_k \in F \tag{3}$$

Where $\hat{y}_i$ is the predicted output, f_k are the weak learners, and F is the space of all possible trees.

Random Forest (RF)

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T(x, \theta_i) \tag{4}$$

Where T represents the decision trees, θ_i are the parameters of each tree, and N is the number of trees.

GBDT

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i), \ f_k \in F \tag{5}$$

Where F is the space of all possible trees.

3.6 Validation and Performance Assessment

The final phase involves validation and performance assessment through rigorous five-fold cross-validation, featuring systematic data partitioning and comprehensive performance evaluation. Multiple performance metrics were calculated, including Mean Absolute Error (MAE) for error magnitude assessment, Root Mean Square Error (RMSE) for prediction accuracy evaluation, Mean Absolute Percentage Error (MAPE) for scale-independent error measurement, and R Square (R^2) and Adjusted R Square for model fit assessment. The methodology concludes with in-depth feature importance analysis and systematic results documentation, providing valuable insights for healthcare planning and policy development.

3.6.1 Cross Validation Strategy

Following Hu et al.'s (2020) method, a five-fold cross-validation strategy was implemented using KFold for robust model evaluation. The cross-validation process comprised several key components. First, data splitting divided the dataset into training and testing sets across five folds. SFFS was then employed for dimensionality reduction and feature selection within each fold. The model underwent training on the designated training set, followed by performance evaluation on the

held-out testing set using comprehensive metrics including MAE, RMSE, MAPE, R^2, and Adjusted R^2. For models supporting feature importance estimation, these values were computed to enable further analysis. To ensure reproducibility and future reference, all trained models, feature selection objects, and feature importance calculations were systematically saved.

3.6.2 Performance Metrics

The evaluation methodology, adopted from Hu et al. (2020), employed five key performance metrics:

Mean Absolute Error (MAE): The average absolute difference between predictions and actual values - e.g., an MAE of 5 means predictions are off by 5 units on average.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

where n is the number of samples, y_i are the true values, and $\hat{y}_i$ are the predicted values.

Root Mean Square Error (RMSE): The average prediction error with higher weight on large deviations - e.g., an RMSE of 7 means most predictions fall within 7 units of actual values, with larger errors having more impact.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (7)$$

where n is the number of samples, y_i are the true values, and $\hat{y}_i$ are the predicted values.

Mean Absolute Percentage Error (MAPE): The average percentage difference between predictions and actual values - e.g., a MAPE of 15% means predictions are off by 15% on average.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \quad (8)$$

Where, y_i represents the actual value for observation i, $\hat{y}_i$represents the predicted value for observation i, n is the total number of observations.

R Square: The proportion of variance explained by the model - e.g., an R^2 of 0.75 means the model explains 75% of the variation in the data.

$$R^2 = 1 - \frac{\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i-1}^{n}(y_i - \bar{y})^2} \quad (9)$$

Additionally, adjusted R square was also added in this study.

Adjusted R Square: A statistical metric that modifies accuracy based on the number of predictors - e.g., if adding a variable only changes the Adjusted R Square from 0.75 to 0.751, that variable may not improve model performance meaningfully.

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{(n-p-1)} \quad (10)$$

Where R^2 is the ordinary coefficient of determination, n is the number of observations, p is the number of predictors in the model ($p = Number\ of\ SDHs\ features + 1$).

This chapter detailed the methodological framework used to predict NCD prevalence using SDH-related features. The approach integrated multiple components: development tools, feature selection methods, various algorithms in a stacking model design, and comprehensive evaluation metrics. This systematic methodology, based on Hu et al. (2020)'s successful approach, was designed to ensure reliable and reproducible results in predicting NCD prevalence patterns in Thailand. The next chapter presents the implementation results and analysis of this methodology.

CHAPTER 4

FINDINGS

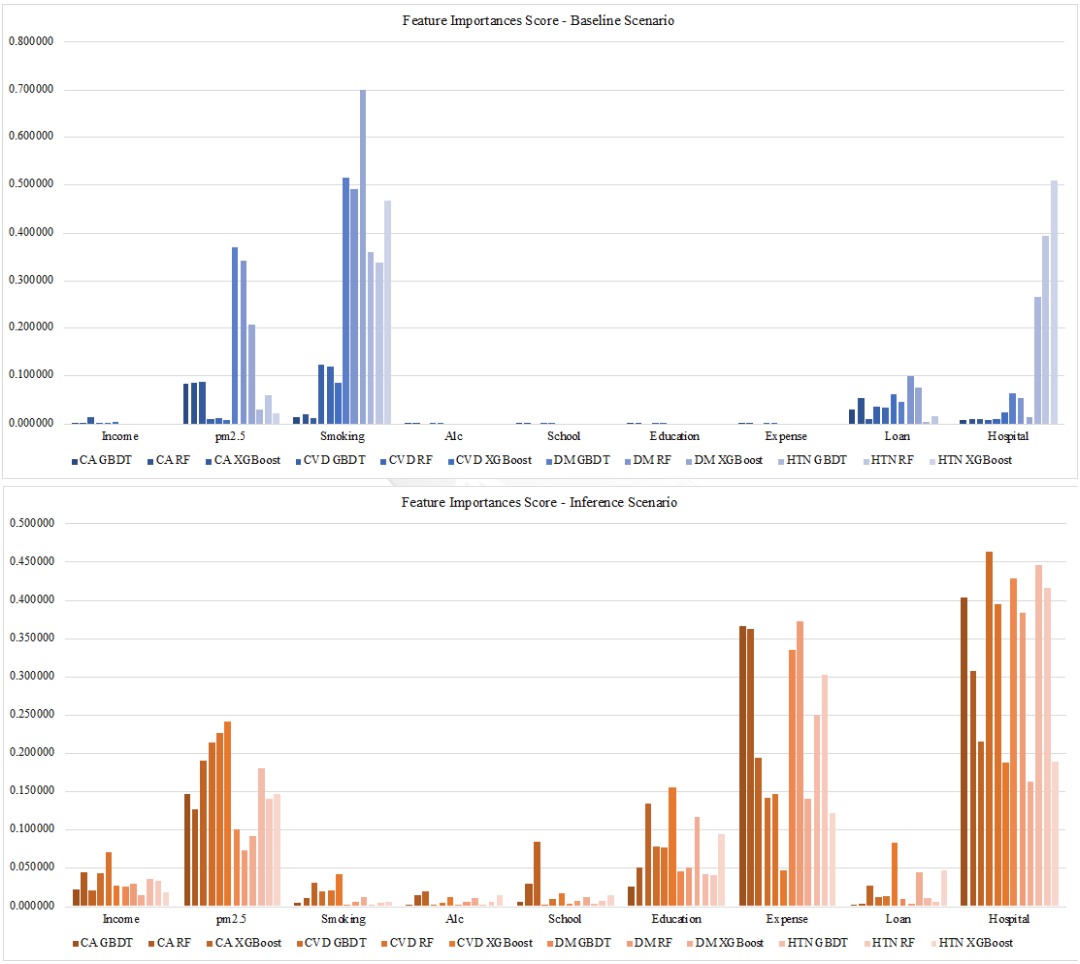
In this chapter, the normalization process was undertaken, illuminating its critical role in data preparation for analysis. Through meticulous examination and analysis, the report offered insights into both the predictive outcomes and the foundational data preprocessing techniques employed. Subsequently, the results obtained from the implemented predictive framework were presented.

4.1 Feature Importance Analysis

Understanding the factors that underlie model predictions is crucial in predictive modelling, particularly in NCD. This study conducted a comparative analysis of feature importance scores derived from both baseline and inference models using the provided dataset, as illustrated in Figure 4.1 and Table 4.1. The included features were including household income, expenses, loans, years of education, number of educational institutions, number of hospitals in each province, number of smokers, concentration of pm2.5, and number of alcoholic consumers.

The analysis investigated the shifts in feature importance scores across different diseases and models under baseline and inference conditions. Within each disease-model combination (e.g., CVD with Gradient Boosted Decision Trees [GBDT]), comparisons were made between baseline and inference scores, revealing fluctuations. Examining diseases within the same model and experimental setup (e.g., CVD vs. Stroke under baseline conditions using GBDT) shed light on how feature importance differed. This comparative analysis facilitated an understanding of how individual features contributed variably to the prediction of different diseases.

Figure 4.1: Feature Importance Score in Visualisation



CA, CVD, DM, HTN	Noncommunicable diseases (NCD)
GBDT, RF, XGBoost	Algorithms

In Figure 4.1, the overview revealed distinct patterns in feature importance across different scenarios. In the baseline scenario, the analysis indicated that variables such as the number of alcohol consumers, years of education, number of educational institutions, household expenses, and loans exerted minimal influence on the experimental models, while the number of smokers emerged as the most prominent factor. Conversely, in the inference scenario, the number of hospitals in each province garnered the highest score, suggesting its pivotal role in the predictive models. Conversely, variables like the number of smokers, alcohol consumers, household loans, and educational institutions received lower scores in this scenario.

These findings underscored the importance of contextual analysis in understanding model behavior and outcomes. In the baseline scenario, where individual behaviors like smoking habits held greater sway, societal indicators such as education and household expenses appeared to have limited predictive power. On the other hand, the inference scenario prioritized the availability and distribution of healthcare resources, as indicated by the prominence of the number of hospitals in each province.

4.1.1 Feature Importance in CA

In CA dataset, it seemed to favour in inference scenario for those three of models. The notable detail as follow:

4.1.1.1 GBDT model

In the inference scenario, features like household income, concentration of pm2.5 (air pollution), years of education, number of educational institutions, household expense, and number of hospitals in each province showed higher importance compared to the baseline. Other features had lower importance scores in the inference scenario.

4.1.1.2 RF model

Similar to GBDT, in the inference scenario, household income, concentration of pm2.5 (air pollution), number of alcoholic consumers, years of education, number of educational institutions, household expense, and number of hospitals in each province exhibited higher importance. Other features had lower scores in the baseline scenario.

4.1.1.3 XGBoost model

All features had higher scores in the inference scenario compared to the baseline.

4.1.2 Feature Importance in CVD

In CVD dataset, similarly to CA dataset, it seemed to favour in inference scenario for those three of models. The notable detail as follow:

4.1.2.1 GBDT model

In the inference scenario, features like household income, concentration of pm2.5 (air pollution), years of education, number of educational institutions, household expense, and number of hospitals in each province showed higher

importance compared to the baseline. Similar to GBDT model in CA dataset. Other features had lower importance scores in the inference scenario.

4.1.2.2 RF model

Similar to GBDT, in the inference scenario, household income, concentration of pm2.5 (air pollution), number of alcoholic consumers, years of education, number of educational institutions, household expense, and number of hospitals in each province exhibited higher importance. Similar to RF model in CA dataset. Other features displayed lower importance scores in the inference scenario compared to the baseline.

4.1.2.3 XGBoost model

All features had higher importance scores in the inference scenario compared to the baseline, except for number of smokers.

4.1.3 Feature Importance in DM

In DM dataset, similarly to CA and CVD datasets, it seemed to favour in inference scenario for those three of models. The notable detail as follow:

4.1.3.1 GBDT model

Household income, years of education, number of educational institutions, household expense, number of alcoholic consumers, and number of hospitals in each province showed increased importance in the inference scenario. Other factors displayed lower importance scores compared to the baseline.

4.1.3.2 RF model

Similar to GBDT, household income, years of education, number of educational institutions, household expense, number of alcoholic consumers, and number of hospitals in each province showed higher importance in the inference scenario. Other factors had lower importance scores compared to the baseline similar to GBDT model.

4.1.3.3 XGBoost model

Similar to GBDT and RF models, household income, years of education, number of educational institutions, household expense, number of alcoholic consumers, and number of hospitals in each province showed higher importance in the inference scenario. Other factors had lower importance scores compared to the baseline similar to GBDT and RF models.

4.1.4 Feature Importance in HTN

In HTN dataset, similarly to previous datasets, it seemed to favour in inference scenario for those three of models. The notable detail as follow:

4.1.4.1 GBDT model

All features showed increased importance in the inference scenario, except for number of smokers.

4.1.4.2 RF model

All features exhibited higher importance in the inference scenario, except for number of smokers and household loan.

4.1.4.3 XGBoost model

All features had higher importance scores in the inference scenario, except for number of smokers and number of hospitals in each province.

4.1.5 Feature Importance in COPD

In COPD dataset, similarly to previous datasets, it seemed to favour in inference scenario for those three of models. The notable detail as follow:

4.1.5.1 GBDT model

All features exhibited higher importance in the inference scenario, except for number of smokers and household loan.

4.1.5.2 RF model

Similar to GBDT model, all features exhibited higher importance in the inference scenario, except for number of smokers and household loan.

4.1.5.3 XGBoost model

All features had higher importance scores in the inference scenario compared to the baseline, except for number of smokers.

4.1.6 Feature Importance in Stroke

In stroke dataset, similarly to previous datasets, it seemed to favour in inference scenario for those three of models. The notable detail as follow:

4.1.6.1 GBDT model

Similar to GBDT model in COPD dataset, all features exhibited higher importance in the inference scenario, except for number of smokers and household loan.

4.1.6.2 RF model

Similar to GBDT model, all features exhibited higher importance in the inference scenario, except for number of smokers and household loan.

4.1.6.3 XGBoost model

All features had higher importance scores in the inference scenario compared to the baseline, except for number of smokers.

The feature importance analysis across various datasets (CA, CVD, DM, HTN, COPD, Stroke) and models (GBDT, RF, XGBoost), as detailed in Table 4.1, consistently highlighted a preference for the inference scenario. In the CA dataset, the GBDT, RF, and XGBoost models all showed higher importance for features like household income, pm2.5 concentration, education-related variables, household expenses, and hospital counts during inference. Similar trends were observed in the CVD and DM datasets. In HTN and COPD datasets, all models favored features except for smokers and household loans. However, in the Stroke dataset, while all models showed increased importance in the inference scenario, the exclusion of smokers was consistent. Overall, these findings suggest a consistent pattern across datasets and models where certain key features emerge as contributors to the inference scenario, providing insights into the predictive capabilities of the models and the factors influencing the specific NCD.

Table 4.1: Feature Importance Score

NCD	Scenario	Model	Feature Importance Score								
			Income	pm2.5	Smoking	Alcohol	School	Education	Expense	Loan	Hospital
CA	Baseline	GBDT	0.002	0.083	0.0156	0.0026	0.0027	0.0025	0.0027	0.0303	0.0065
	Inference	GBDT	0.0218	0.1462	0.0042	0.0018	0.0054	0.0255	0.367	0.0009	0.4042
	Baseline	RF	0.002	0.0862	0.0183	0.0013	0.0024	0.0014	0.003	0.0543	0.0098
	Inference	RF	0.0448	0.1273	0.0101	0.0155	0.03	0.0505	0.3629	0.0027	0.3084
	Baseline	XGBoost	0.0142	0.0878	0.0126	0	0	0	0	0.0101	0.01
	Inference	XGBoost	0.021	0.1903	0.0315	0.0198	***0.0846***	0.1346	0.1936	0.0267	0.2164
CVD	Baseline	GBDT	0.001	0.0107	0.1237	0.0007	0.0011	0.0006	0.0012	0.0359	0.0066
	Inference	GBDT	0.0424	0.214	0.0194	0.0007	0.0016	0.0785	0.1416	0.0123	0.4646
	Baseline	RF	0.0026	0.0125	0.1207	0.0012	0.002	0.001	0.0006	0.0341	0.0095
	Inference	RF	***0.071***	0.227	0.0208	0.0049	0.0089	0.0772	0.1465	0.0138	0.3954
	Baseline	XGBoost	0.0042	0.0066	0.0863	0	0	0	0	0.0626	0.023
	Inference	XGBoost	0.0261	0.2413	0.0414	0.0115	0.0161	***0.1562***	0.0472	0.0835	0.1877
DM	Baseline	GBDT	0	***0.3696***	0.517	0	0	0	0	0.0447	0.0632
	Inference	GBDT	0.0252	0.1013	0.0009	0.0005	0.0038	0.0456	0.3363	0.0089	0.4289
	Baseline	RF	0	0.3405	0.4917	0	0	0	0	***0.1005***	0.0547
	Inference	RF	0.0298	0.0732	0.0068	0.0057	0.0073	0.0505	***0.3726***	0.0024	0.3842
	Baseline	XGBoost	0	0.2092	***0.7005***	0	0	0	0	0.0749	0.0153
	Inference	XGBoost	0.0154	0.0913	0.0113	0.0103	0.0122	0.1176	0.14	0.0444	0.1625

(Continued)

Table 4.1 (Continued): Feature Importance Score

NCD	Scenario	Model	Feature Importance Score								
			Income	**pm2.5**	**Smoking**	**Alcohol**	**School**	**Education**	**Expense**	**Loan**	**Hospital**
HTN	Baseline	GBDT	0	0.0316	0.3603	0	0	0	0	0.0042	0.2659
	Inference	GBDT	0.0353	0.1806	0.0002	0.0007	0.0034	0.0412	0.2506	0.0109	0.4471
	Baseline	RF	0	0.0584	0.3388	0	0	0	0	0.0168	0.3932
	Inference	RF	0.033	0.1401	0.0046	0.0058	0.0082	0.0404	0.303	0.0056	0.4173
	Baseline	XGBoost	0	0.0217	0.4685	0	0	0	0	0	0.5098
	Inference	XGBoost	0.0176	0.1467	0.0066	0.0156	0.0143	0.0955	0.1214	0.048	0.189
COPD	Baseline	GBDT	0.0003	0.054	0.2384	0.0004	0.0008	0.0007	0.0008	0.0478	0.3877
	Inference	GBDT	0.0201	0.1806	0.0007	0.0059	0.0012	0.0088	0.1648	0.0013	***0.6075***
	Baseline	RF	0.0008	0.0278	0.2094	0.0012	0.0007	0.001	0.0012	0.0368	0.3176
	Inference	RF	0.0223	0.173	0.005	0.0089	0.0073	0.0213	0.1737	0.0006	0.5672
	Baseline	XGBoost	0.0024	0.0326	0.2256	0	0	0	0	0	0.2571
	Inference	XGBoost	0.0262	0.2979	0.0262	***0.0488***	0.0199	0.0625	0.1193	0.0388	0.3604
Stroke	Baseline	GBDT	0.001	0.0104	0.1237	0.0008	0.0013	0.0013	0.0008	0.0354	0.0077
	Inference	GBDT	0.0244	0.2421	0.001	0.001	0.0024	0.0582	0.1427	0.0096	0.4847
	Baseline	RF	0.0016	0.0193	0.1129	0.0024	0.0021	0.0014	0.0032	0.0297	0.0103
	Inference	RF	0.0532	0.2312	0.0078	0.0057	0.0115	0.0552	0.1375	0.0118	0.4295
	Baseline	XGBoost	0.0042	0.0066	0.0863	0	0	0	0	0.0626	0.023
	Inference	XGBoost	0.0199	0.2036	0.01	0.0258	0.0173	0.1178	0.0515	0.0643	0.2155

*Bold and Italic numbers = Highest value within that particular feature

4.2 Model Evaluation

This section analysed the performance of various machine learning models for predicting the outcomes of different NCD including CA, DM, CVD, HTN, COPD, and Stroke as demonstrated in Table 4.2. The models' effectiveness was compared in two scenarios:

4.2.1 Baseline Scenario

Handling Missing Values: In the baseline scenario, rows containing missing values were removed from the dataset. This technique is known as complete-case deletion and is considered a simple but potentially problematic approach. It can lead to biased results by discarding potentially valuable data points.

No Imputation: No imputation was applied to missing values in the baseline scenario. This means the models were trained and evaluated directly on the remaining data points without any attempt to fill in the missing information.

4.2.2 Inference Scenario

Handling Missing Values: During the inference scenario, missing values were imputed using the average value for each feature (column) in the dataset. This is a basic imputation technique called mean imputation. While straightforward, it can be problematic if the average value doesn't accurately represent the missing data, potentially introducing bias.

Applying the Models: The trained models were then applied to predict outcomes on new, unseen data that may also contain missing values. These missing values were filled in using the calculated average values obtained during the previous step.

To provide a comprehensive overview of all model performances across different NCDs, Table 4.2 presents the comparison of evaluation metrics for both baseline and inference scenarios.

Table 4.2: Summary of Comparison of Model Performance Metrics Across Different NCDs in Baseline and Inference Scenarios

Dataset	Scenario	Model	MAE	RMSE	MAPE	R^2	Adj. R^2
CA	Baseline	SVR	42.11	-	115.77	(neg)	(neg)
		GBDT/LR/RF/XGB	43.97 to 44.03	74.92 to 75.04	-	(0.02) to 0.04	(0.02) to 0.04
	Inference	SVR	36.21	75.93	133.14	0.01	0.01
		GBDT/LR/RF/Stack/XGB	35.46 to 39.22	71.89 to 75.12	-	0.03 to 0.11	0.03 to 0.11
DM	Baseline	SVR	3.67	-	88.78	0.01	0
		GBDT/LR/RF/Stack/XGB	4.08 to 4.18	8.88 to 8.97	160.12 to 166.07	0.06 to 0.08	0.06 to 0.08
	Inference	SVR	2.58	6.74	69.32	-0.03	-0.03
		GBDT/LR/RF/Stack/XGB	2.81 to 3.01	6.37 to 6.59	123.26 to 137.13	0.01 to 0.07	0.01 to 0.07
CVD	Baseline	SVR	0.77	1.44	48.48	0.02	0.02
		GBDT/LR/Stack	0.77	1.44	48.35 to 48.42	0.02	0.02
		RF	0.62	1.53	23.74	-0.11	-0.12
		XGBoost	0.78	1.44	49.03	0.02	0.01
	Inference	SVR	0.77	1.44	48.28	0.02	0.02
		GBDT/LR/Stack	0.73 to 0.74	1.67	45.91 to 46.78	0.03	0.03
		RF	0.58	1.75	22.7	-0.07	-0.07
		XGBoost	0.76	1.69	48.19	0.01	0.01
HTN	Baseline	SVR	19.61	-	95.06	-0.04	-0.05
		GBDT/LR/RF/Stack/XGB	20.74 to 20.77	35.73 to 35.75	120.97 to 121.70	0	0
	Inference	SVR	17.8	37.98	107.36	-0.01	-0.01
		GBDT/LR/RF/Stack/XGB	17.84 to 18.80	36.21 to 37.27	131.49 to 137.34	0.02 to 0.08	0.02 to 0.08

(Continued)

Table 4.2 (Continued): Summary of Comparison of Model Performance Metrics Across Different NCDs in Baseline and Inference Scenarios

Dataset	Scenario	Model	MAE	RMSE	MAPE	R^2	Adj. R^2
COPD	Baseline	SVR	3.67	-	88.78	0.01	0
		GBDT/LR/RF/Stack/XGB	4.08 to 4.18	8.88 to 8.97	160.17 to 166.07	0.06 to 0.08	0.06 to 0.08
	Inference	SVR	2.56	6.63	68.7	-0.04	-0.04
		GBDT/LR/RF/Stack/XGB	2.79 to 2.99	6.27 to 6.48	122.79 to 136.63	0.01 to 0.08	0.01 to 0.08
Stroke	Baseline	SVR	2.05	-	68.87	0.04	0.04
		GBDT/LR/RF/Stack/XGB	2.19 to 2.30	3.42 to 3.52	95.51 to 102.90	0.06 to 0.12	0.06 to 0.12
	Inference	SVR	1.81	3.47	62.37	-0.04	-0.04
		GBDT/LR/RF/Stack/XGB	1.87 to 2.02	3.22 to 3.40	87.95 to 98.47	0.00 to 0.10	0.00 to 0.10

4.2.3 The Result Description for Baseline Scenario

The baseline scenario revealed distinct patterns of model performance across datasets. In the CA dataset, SVR consistently achieved the lowest MAE at 42.11, indicating strong predictive accuracy. However, this model struggled to explain the variability in the data, as shown by its negative R^2 and Adjusted R^2 values. Similarly, in the DM dataset, SVR achieved the lowest MAE at 3.67 but exhibited the highest MAPE (88.78), suggesting its predictions lacked stability for cases with smaller values or outliers.

Ensemble models, including GBDT, RF, and Stacking, demonstrated comparable MAEs but offered better explanatory power. For instance, in the CVD dataset, RF achieved the lowest MAE of 0.62 and MAPE of 23.74, highlighting its capacity to reduce prediction errors while maintaining consistency across different data distributions. However, even with its superior MAE, RF exhibited negative R^2 values (-0.11 to -0.12), indicating poor performance in explaining variance. The HTN dataset presented similar results, where SVR achieved the lowest MAE (19.61) but recorded the highest MAPE (95.06), suggesting difficulty in handling outliers or extreme values.

While SVR demonstrated strong predictive accuracy across datasets, its inability to explain data variance, as evidenced by its low or negative R^2 values, limited its applicability. Ensemble models, such as RF and Stacking, offered a more balanced approach, demonstrating moderate R^2 values while maintaining competitive MAEs, particularly in datasets with higher variability.

4.2.4 The Result Description for Inference Scenario

The inference scenario provided further insights into model robustness and generalizability. For the CA dataset, RF and Stacking consistently achieved the lowest MAEs, ranging from 35.46 to 39.22, with moderate R^2 values (0.03 to 0.11). These findings suggested their ability to generalize well to unseen data while balancing prediction accuracy and explanatory power. By contrast, SVR, despite achieving an MAE of 36.21, continued to exhibit limited explanatory power, with an R^2 value of 0.01.

In the COPD dataset, RF demonstrated its robustness by achieving the lowest MAE (2.56) and a moderate R^2 value (0.01 to 0.08). However, SVR, although achieving competitive MAEs, exhibited negative R^2 values, suggesting that its predictions were not well-aligned with the underlying data structure. Similarly, in the Stroke dataset, SVR achieved the lowest MAE (1.81) and the lowest MAPE (62.37), indicating strong predictive accuracy. Yet, ensemble models like Stacking and XGBoost outperformed SVR in terms of explanatory power, achieving R^2 values ranging from 0.00 to 0.10, while maintaining comparable MAEs.

Notably, in the HTN dataset, RF and Stacking performed exceptionally well, balancing low MAEs (17.84 to 18.80) with moderate R^2 values (0.02 to 0.08). This indicated their ability to effectively manage datasets with higher variability or complex relationships. SVR, although achieving the lowest MAE (17.80), continued to struggle with explaining variance, as shown by its negative R^2 value.

4.2.5 Findings Summary

The findings revealed a consistent trade-off between predictive accuracy and explanatory power, with model performance varying depending on the dataset and scenario. SVR consistently minimized MAEs, particularly in datasets such as CA and DM, where its predictive accuracy was unmatched. However, its limited ability to explain variance, reflected in low or negative R^2 values, restricted its applicability for tasks requiring interpretability. Conversely, ensemble models such as RF, Stacking, and XGBoost exhibited a more balanced performance, achieving moderate R^2 values while maintaining competitive accuracy across all datasets.

4.2.6 Detailed Dataset-Specific Insights

CA Dataset: Ensemble models, particularly RF and Stacking, excelled in both baseline and inference scenarios, demonstrating their capacity to reduce errors while providing moderate explanatory power.

DM Dataset: The SVR model minimized prediction errors but exhibited high MAPE and low R^2 values, indicating challenges in managing datasets with skewed distributions.

CVD Dataset: RF consistently demonstrated superior predictive accuracy (low MAE and MAPE), but its explanatory power was limited.

COPD Dataset: Ensemble models outperformed SVR in inference scenarios, with RF achieving a balance between accuracy and moderate R^2 values.

Stroke Dataset: Stacking and XGBoost emerged as the most balanced models, combining competitive MAEs with superior explanatory power.

HTN Dataset: RF and Stacking demonstrated strong performance across scenarios, highlighting their suitability for datasets with high variability.

These findings provided valuable insights into the strengths and limitations of the predictive models across different scenarios and datasets. Building on this analysis, the next section delved deeper into the broader implications of these results, explored the limitations of the study, proposed strategies for addressing these challenges, and identified opportunities for future research to enhance both predictive accuracy and explanatory power.

CHAPTER 5

DISCUSSION

5.1 Conclusion

The study aimed to predict NCD prevalence using selected SDH-related features and to evaluate the models developed from these features. The analysis of feature importance across various datasets and models provided valuable insights into the predictive capabilities of SDH-related features.

Throughout the study, specific socio-economic and environmental factors, such as household income, air pollution levels, education-related variables, household expenses, and healthcare infrastructure, were identified as significant contributors to predicting the occurrence or progression of NCD. GBDT, RF, and XGBoost model consistently favored the inference scenario, demonstrating the effectiveness of SDH-related features in NCD prevalence prediction.

SVR occasionally exhibited lower MAE in the baseline scenario, but its poor explanatory power highlighted the importance of models with better interpretability, such as GBDT, RF, and XGBoost. During the inference scenario, RF, Stacking, and XGBoost models showcased superior predictive accuracy with lower MAE and RMSE, indicating their potential for minimising prediction errors.

The study has contributed the analysis of feature importance and model performance. The insights gleaned from this study offer valuable guidance for healthcare practitioners and policymakers in devising evidence-based strategies to mitigate the impact of NCD on public health. Further refinement of models may enhance their interpretability and aid in the development of targeted interventions and preventive strategies.

5.2 Discussion

The findings from this study revealed several noteworthy patterns and relationships between SDH and NCD prevalence in Thailand, offering valuable insights into the predictive capabilities of a stacking ensemble method across different disease categories.

The feature importance analysis consistently highlighted household income, air pollution levels (pm2.5), education-related variables, household expenses, and healthcare infrastructure as significant contributors to NCD prediction. This resonates with Stringhini et al.'s (2018) findings on the association between socioeconomic status and physical functioning, while extending the scope to include a broader arrayof SDH domains. Similarly, our identification of pm2.5 as a high-importance feature corresponds with George and Thomas's (2018) emphasis on environmental factors in predicting respiratory disease patterns.

The stacking ensemble methodology, adapted from Hu et al. (2020), demonstrated mixed results across different NCD categories. While our models showed lower explanatory power (with maximum R^2 values of 0.12) compared to Hu et al.'s reported outcomes, our findings nonetheless validate their approach of combining multiple base learners to enhance prediction accuracy. The consistently superior performance of RF across multiple NCD categories aligns with Alim et al.'s (2020) discovery of RF's effectiveness in classifying cardiovascular conditions, suggesting this algorithm possesses robust capabilities for handling the complex, multidimensional relationships between SDH and health outcomes.

Returning to the first research question regarding the effectiveness of stacking ensemble methodologies for predicting NCD prevalence using SDH features, our results indicate modest predictive capabilities across different disease categories. The models demonstrated varying degrees of accuracy, with SVR occasionally achieving lower MAE values but suffering from poor explanatory power, while ensemble models like RF, GBDT, and XGBoost offered a more balanced performance profile. This suggested that while SDH features indeed contribute meaningful information to NCD prediction, their predictive power may be enhanced through integration with traditional clinical risk factors, as explored by Davagdorj et al. (2021).

The second research question sought to identify which specific SDH features demonstrated the highest predictive importance for different NCD categories. The comprehensive feature importance analysis revealed distinct patterns for each disease category, with household income, pm2.5 levels, education-related variables, and healthcare infrastructure consistently emerging as significant contributors. Notably, these findings extended beyond Wang and Wang's (2020) focus on country-level

socioeconomic factors by demonstrating the importance of smaller level SDH features in predicting NCD outcomes within a single nation. The emergence of healthcare infrastructure (hospital counts) as a top feature in several models corresponds with Hastings et al.'s (2022) emphasis on healthcare access as a critical determinant of cardiovascular disease risk across socioeconomic groups.

Regarding the third research question on the impact of different data preprocessing strategies, the findings clearly demonstrate that the inference scenario (with mean imputation) generally yielded superior results compared to the baseline scenario (with complete case analysis). This pattern was consistent across different NCD categories and model types, suggesting that preserving data points through imputation provides more robust training examples for the models. The superior performance of the inference scenario underscores the importance of addressing missing values effectively when working with complex, real-world SDH datasets characterized by incomplete coverage across provinces or hospitals and time periods.

In examining the specific performance metrics across different NCD categories, several patterns emerged that warrant further discussion. For instance, the models demonstrated better predictive performance for Stroke and COPD compared to HTN and DM, as evidenced by higher R^2 values. This variability might reflect differences in how strongly these conditions are influenced by the particular SDH factors included in our dataset, or it could indicate that certain diseases exhibit more discernible patterns in relation to social determinants. These findings aligned with Nawamawat et al.'s (2019) observation that different NCD have varying sensitivity to socioeconomic and environmental factors in Thailand.

The consistently strong performance of ensemble methods, particularly RF and Stacking, across multiple disease categories suggests that these approached are well-suited to capturing the complex, non-linear relationships between SDH and NCD prevalence. This aligned with Hu et al.'s (2018) demonstration of ensemble methods' effectiveness in predicting non-communicable diseases in Bangladesh, suggesting the transferability of these methodological approaches across different Southeast Asian contexts.

While the feature importance analysis yielded valuable insights, it's worth noting the substantial shift in feature importance patterns between baseline and inference scenarios. This shift underscores the sensitivity of machine learning models to preprocessing decisions and highlights the need for careful consideration when interpreting feature importance in the context of SDH-based NCD prediction. The variation in feature importance across different disease categories further suggests that the social determinants of health may exert disease-specific influences, rather than affecting all NCD uniformly - a nuance that previous studies such as Potempa et al. (2022) have acknowledged but not extensively quantified.

5.3 Limitation

The study encountered several significant constraints that influenced the analytical scope and results. Data consistency represented a primary challenge, with noticeable variability in the availability and completeness of time-series data across different SDH features at the provincial level. Measurements for indicators such as smoking rates, alcohol consumption patterns, and educational attainment lacked uniform collection methods or regular intervals, potentially introducing biases in the relationships established between these features and NCD prevalence. This irregularity in data collection hampered comprehensive temporal analysis and limited the ability to accurately capture how social determinants evolved alongside disease patterns over time.

Air pollution monitoring presented another substantial limitation, with pm2.5 measurements available only for certain regions and time periods. The incomplete spatial coverage of air quality data restricted the analysis of how environmental factors contributed to health outcomes across Thailand's diverse geographic landscapes. This constraint was particularly significant given that the feature importance analysis consistently identified pm2.5 levels as influential predictors for multiple NCD categories.

The imbalanced explanatory power of the models, evidenced by modest R-squared values (maximum of 0.12), indicated underlying issues with either feature selection or model architecture. These results suggested that while the selected SDH

features provided meaningful predictive information, they captured only a portion of the complex factors determining NCD prevalence patterns. The models exhibited varying performance across different disease categories, with certain conditions showing greater predictability than others, pointing to disease-specific relationships with the selected determinants.

Methodological limitations also emerged during the model evaluation process. The substantial differences in feature importance rankings between baseline and inference scenarios highlighted the sensitivity of machine learning approaches to data preprocessing decisions. This sensitivity complicated the interpretation of which SDH features truly exerted the strongest influence on NCD outcomes, as the apparent importance shifted depending on how missing values were handled.

The absence of individual-level data alongside population-level SDH metrics created another analytical gap. Without this integration, the models could not account for how individual risk factors interacted with broader social determinants, potentially overlooking important mechanisms through which SDH influenced disease development at the personal level. This restriction limited the ability to distinguish between population-level effects and individual susceptibility patterns.

Finally, while the stacking ensemble methodology offered improved performance over single models in several instances, the overall predictive capability remained less robust than anticipated based on previous research such as Hu et al. (2020). This discrepancy suggested that additional factors beyond those included in the current feature set may play significant roles in determining NCD prevalence in Thailand, pointing to opportunities for model refinement in future research.

5.4 Future Work

Future research endeavors should build upon the findings and limitations of this study to enhance understanding of the relationships between SDH and NCD prevalence in Thailand. Integration of traditional clinical and SDH features represents a promising direction for further investigation, where hybrid models could combine individual-level clinical data with population-level SDH features. This integrated approach would bridge the gap between clinical risk assessment and social

determinant analysis, providing a more holistic understanding of NCD development pathways across different population segments.

Temporal analysis of SDH-NCD relationships deserved substantial attention in subsequent studies. Longitudinal research tracking how changes in specific SDH features correlate with subsequent shifts in NCD prevalence patterns over time would enable researchers to distinguish between immediate and delayed effects of social determinants on health outcomes. Such temporal perspectives might reveal causal pathways not detectable in cross-sectional analyses and help anticipate future disease burden based on current social trends.

The exploration of regional variations in feature importance constituted another valuable research direction. More granular analysis examining how the predictive importance of specific SDH features varies across different geographic regions within Thailand could reveal regionally-specific determinants of health. These insights would guide the development of localized intervention strategies tailored to each area's unique social context, potentially improving resource allocation efficiency in public health initiatives.

Given the observed variations in model performance across different disease categories, development of specialized models for specific NCD categories warrants further exploration. Future work could focus on developing specialized prediction frameworks optimized for each major NCD type, incorporating disease-specific SDH features and model architectures selected to match each condition's unique etiology and progression patterns. This targeted approach might yield more accurate predictions than generalized models attempting to address all NCD categories simultaneously.

Investigation of interaction effects between SDH domains represented an underexplored area with significant potential. More sophisticated modeling techniques could capture the complex interactions between different domains of social determinants. For instance, examining how educational attainment interacts with economic stability to influence NCD outcomes beyond their individual effects might reveal synergistic or antagonistic relationships between determinants that current models fail to capture.

The assessment of intervention effectiveness through predictive modeling could translate research findings into practical public health applications. Future studies could leverage predictive frameworks developed in this research to evaluate the potential impact of public health interventions targeting specific social determinants. By simulating changes in key SDH features, researchers could estimate the expected effects on NCD prevalence and identify the most promising intervention targets for maximizing public health benefit.

Enhancement of data preprocessing techniques remains crucial for improving model performance with incomplete datasets. Building on findings regarding the impact of different preprocessing strategies, future work could explore more sophisticated approaches to handling missing values in SDH datasets. Methods such as multiple imputation or advanced machine learning-based imputation techniques could potentially further improve model performance beyond the mean imputation approach utilized in this study, addressing one of the fundamental challenges in working with real-world public health data.

BIBLIOGRAPHY

Ai, X. X., Jia, H., & Xin, L. (2016). SVM-based Cancer Incidence Forecasting of Patients. 2016 9th *International Symposium on Computational Intelligence and Design (ISCID).* https://doi.org/10.1109/iscid.2016.2074

Alim, M. A., Habib, S., Farooq, Y., & Rafay, A. (2020). Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble Learning Model. 2020 3rd *International Conference on Computing, Mathematics and Engineering Technologies (ICoMET).* https://doi.org/10.1109/icomet48670.2020.9074135

Banu, M. A. N., & Gomathy, B. (2014). Disease Forecasting System Using Data Mining Methods. 2014 *International Conference on Intelligent Computing Applications*. https://doi.org/10.1109/icica.2014.36

Bhandari, A. (2024, January 16). Key Difference between R-squared and Adjusted R-squared for Regression Analysis. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2020/07/difference-between-r-squared-and-adjusted-r-squared/

Bhandari, P. (2022a, November 11). How to Find Outliers | 4 Ways with Examples & Explanation. *Scribbr*. https://www.scribbr.com/statistics/outliers/

Bhandari, P. (2022b, November 11). Missing Data | Types, Explanation, & Imputation. *Scribbr*. https://www.scribbr.com/statistics/missing-data/

Bhoothookngoen, P., & Sanchan , N. (2024). Prevalence of Noncommunicable Diseases and Social Determinants of Health in Thailand: Insights from Public Datasets. *Thai Journal of Public Health*, 54(2).

Bhoothookngoen, P., & Sanchan , N. (2023). Predictive Modeling of Non-Communicable Diseases Using Social Determinants of Health as Features: A Review of Existing Approaches. *Srinakharinwirot University Engineering Journal,* 19(1), 79–88. Retrieved from https://ph02.tci-thaijo.org/index.php/sej/article/view/249846

Chakarverti, M., Yadav, S., & Rajan, R. (2019). Classification Technique for Heart Disease Prediction in Data Mining. 2019 2nd *International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT).* https://doi.org/10.1109/icicict46008.2019.8993191

Cost of Illness | POLARIS | Policy and Strategy | CDC. (n.d.). Retrieved from https://www.cdc.gov/policy/polaris/economics/cost-illness/index.html

Davagdorj, K., Bae, J. W., Pham, V. H., Theera-Umpon, N., & Ryu, K. H. (2021). Explainable Artificial Intelligence Based Framework for Non-Communicable Diseases Prediction. *IEEE Access*, 9, 123672–123688. https://doi.org/10.1109/access.2021.3110336

Davagdorj, K., Pham, V. H., Theera-Umpon, N., & Ryu, K. H. (2020). XGBoost-Based Framework for Smoking-Induced Noncommunicable Disease Prediction. *International Journal of Environmental Research and Public Health*, 17(18), 6513. https://doi.org/10.3390/ijerph17186513

Ferdousi, R., Hossain, M. A., & El Saddik, A. (2021). Early-Stage Risk Prediction of Non-Communicable Disease Using Machine Learning in Health CPS. *IEEE Access*, 9, 96823–96837. https://doi.org/10.1109/access.2021.3094063

GeeksforGeeks. (2024, March 20). Linear Regression in Machine learning. *GeeksforGeeks*. https://www.geeksforgeeks.org/ml-linear-regression/

George, N., & Thomas, J. (2018). Forecasting the Peak Demand Days of Chronic Respiratory Diseases with Fuzzy Logic. *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET).* https://doi.org/10.1109/iccsdet.2018.8821106

Ge, R., Zhang, R., & Wang, P. (2020). Prediction of Chronic Diseases With Multi-Label Neural Network. *IEEE Access*, 8, 138210–138216. https://doi.org/10.1109/access.2020.3011374

Hastings, K., Marquina, C., Morton, J., Abushanab, D., Berkovic, D., Talic, S., Zomer, E., Liew, D., & Ademi, Z. (2022). Projected New-Onset Cardiovascular Disease by Socioeconomic Group in Australia. *PharmacoEconomics*, 40(4), 449–460. https://doi.org/10.1007/s40273-021-01127-1

Hu, M., Nohara, Y., Wakata, Y., Ahmed, A., Nakashima, N., & Nakamura, M. (2018). Machine Learning Based Prediction of Non-communicable Diseases to Improving Intervention Program in Bangladesh. *European Journal for Biomedical Informatics*, 14(4). https://doi.org/10.24105/ejbi.

Hu, Z., Qiu, H., Su, Z., Shen, M., & Chen, Z. (2020). A Stacking Ensemble Model to Predict Daily Number of Hospital Admissions for Cardiovascular Diseases. *IEEE Access*, 8, 138719–138729. https://doi.org/10.1109/access.2020.3012143

Islam, S., Jahan, N., & Khatun, M. E. (2020). Cardiovascular Disease Forecast using Machine Learning Paradigms. 2020 *Fourth International Conference on Computing Methodologies and Communication (ICCMC).* https://doi.org/10.1109/iccmc48092.2020.iccmc-00091

likebupt, PeterCLu, and v-chmccl. (2021, November 4). Normalize Data: Component Reference – Azure Machine Learning. *Microsoft Learn*. https://learn.microsoft.com/en-us/azure/machine-learning/component reference/normalize-data

Kathirvel, S., & Rapporteurs, J. S. T. (2018). Sustainable development goals and noncommunicable diseases: Roadmap till 2030 – A plenary session of world noncommunicable diseases congress 2017. *International Journal of Noncommunicable Diseases*, 3(1), 3. https://doi.org/10.4103/jncd.jncd_1_18

Keerthi Samhitha, B., Sarika Priya., M., Sanjana., C., Mana, S. C., & Jose, J. (2020). Improving the Accuracy in the Prediction of Heart Disease using Machine Learning Algorithms. 2020 *International Conference on Communication and Signal Processing (ICCSP).* https://doi.org/10.1109/iccsp48568.2020.9182303

Ministry of Public Health of Thailand, World Health Organization, United Nations Development Programme, and United Nations Inter-Agency Task Force. (2021). Prevention and control of noncommunicable diseases in Thailand – the case for investment. Retrieved from https://www.who.int/thailand/activities/NCD_Investment_Case_Report

Ministry of Public Health of Thailand. (n.d.). Hospitals Code. Retrieved September 23, 10 C.E., from https://hcode.moph.go.th/code/

Mohan, N., Jain, V., & Agrawal, G. (2021). Heart Disease Prediction Using Supervised Machine Learning Algorithms. 2021 5th *International Conference on Information Systems and Computer Networks (ISCON).* https://doi.org/10.1109/iscon52037.2021.9702314

National Statistical Office Thailand. (n.d.). National Statistical Office Thailand. Retrieved September 22, 10 C.E., from http://www.nso.go.th/

Nawamawat, J., Prasittichok, W., Prompradit, T., Chatchawanteerapong, S., & Sittisart, V. (2020). Prevalence and characteristics of risk factors for non-communicable diseases in semi-urban communities. *Journal of Health Research*, 34(4), 295–303. https://doi.org/10.1108/jhr-03-2019-0058

Ngom, F., Fall, I. S., Camara, M., & Bah, A. (2020). A study on predicting and diagnosing non-communicable diseases: case of cardiovascular diseases. In *International Conference on Intelligent Systems*. https://doi.org/10.1109/iscv49265.2020.9204022

Nipaporn U. (2016). Social Determinants of Health and Health Promotion in Population. Health Technical Office, *Ministry of Public Health, Thailand: Journal of Health Science*, 25(1).

Open Government Data of Thailand. (n.d.). Retrieved from https://opendata.data.go.th/en/group/public-health

Pollution Control Department. (2022, December 14). Air4Thai. http://air4thai.pcd.go.th/webV2/region.php?region=0

Potempa, K., Rajataramya, B., Singha-Dong, N., Furspan, P., Kahle, E., & Stephenson, R. (2022). Thailand's Challenges of Achieving Health Equity in the Era of Non-Communicable Disease. *Pacific Rim international journal of nursing research*, 26(2), 187–197.

Sangkatip, W., & Phuboon-Ob, J. (2020). Non-Communicable Diseases Classification using Multi-Label Learning Techniques. 2020 - 5th *International Conference on Information Technology (InCIT).* https://doi.org/10.1109/incit50588.2020.9310978

Sklearn.ensemble.GradientBoostingRegressor. (n.d.). *Scikit-learn*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

Sklearn.ensemble.RandomForestRegressor. (n.d.). *Scikit-learn*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Sklearn.svm.SVR. (n.d.). *Scikit-learn*. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

Social Determinants of Health - Healthy People 2030 | health.gov. (n.d.). Retrieved from https://health.gov/healthypeople/priority-areas/social-determinants-health

Social determinants of health. (2019, May 30). *World Health Organization*. https://www.who.int/health-topics/social-determinants-of-health

Stringhini, S., Carmeli, C., Jokela, M., Avendaño, M., McCrory, C., d'Errico, A., Bochud, M., Barros, H., Costa, G., Chadeau-Hyam, M., Delpierre, C., Gandini, M., Fraga, S., Goldberg, M., Giles, G. G., Lassale, C., Kenny, R. A., Kelly-Irving, M., Paccaud, F., . . . Kivimäki, M. (2018). Socioeconomic status, non-communicable disease risk factors, and walking speed in older adults: multi-cohort population based study. *BMJ*, k1046. https://doi.org/10.1136/bmj.k1046

THE 17 GOALS | Sustainable Development | United Nations. (n.d.). Retrieved October 15, 2022, from https://SDG.un.org/goals

The Ministry of Public Health of Thailand. (2016, November 1). The Ministry of Public Health of Thailand. Retrieved December 22, 12 C.E., from https://hss.moph.go.th/fileupload_doc_slider/2016-12-01--431.xls

Urwannachotima, N. (2016). Social Determinants of health and health promotion in population. *Journal of Health Science*, 25(1), 147-156.

Wang, Y., & Wang, J. (2020). Modelling and prediction of global non-communicable diseases. *BMC Public Health*, 20(1). https://doi.org/10.1186/s12889-020-08890-4

Warudkar, H. (2022, September 14). How to Find Outliers In Machine Learning: The Guide. *Express Analytics*. https://www.expressanalytics.com/blog/outliers-machine-learning/

What is Normalization in Machine Learning | Deepchecks. (2021, August 5). *Deepchecks*. https://deepchecks.com/glossary/normalization-in-machine-learning/

World Bank. (n.d.). Current health expenditure (% of GDP) - Thailand | Data. *Data.worldbank.org*. Retrieved from https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS?locations=TH

World Health Organization. (2010, July 13). A Conceptual Framework for Action on the Social Determinants of Health. https://www.who.int/publications/i/item/9789241500852.

World Health Organization. Regional Office for Europe. (2016). Action plan for the prevention and control of noncommunicable diseases in the WHO European Region. https://apps.who.int/iris/handle/10665/341522.

World Health Organization. (n.d.). Global Health Estimates: Leading Causes of Death. *World Health Organization*. Retrieved from https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death

World Health Organization. (2022). Non-communicable diseases. (2022, September 16). Retrieved from https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

XGBoost Documentation — xgboost 2.0.3 documentation. (n.d.). https://xgboost.readthedocs.io/en/stable/

APPENDIX

Appendix 1: Publication

Bhoothookngoen, P., & Sanchan , N. (2024). Prevalence of Noncommunicable Diseases and Social Determinants of Health in Thailand: Insights from Public Datasets. Thai Journal of Public Health, 54(2).

Bhoothookngoen, P., & Sanchan , N. (2023). Predictive Modeling of Non-Communicable Diseases Using Social Determinants of Health as Features: A Review of Existing Approaches. Srinakharinwirot University Engineering Journal, 19(1), 79–88. Retrieved from https://ph02.tci-thaijo.org/index.php/sej/article/view/249846

Bhoothookngoen, P., & Sanchan , N. Forecasting Noncommunicable Diseases in Thailand: Evaluating the Predictive Power of Social Determinants of Health-Related Features has been accepted and expected to be published in Srinakharinwirot University Engineering Journal in volume no.21.

Appendix 2: Abbreviations

Abbreviation	Term	Definition
AUC	Area Under the Curve	A metric used to evaluate the performance of a machine learning algorithm, specifically for binary classification problems. AUC represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance, with values ranging from 0 to 1 (higher is better).
CA	Cancer	A group of diseases in this study including Lung Cancer, Cervical Cancer, and Breast Cancer, characterized by abnormal cell growth with potential to invade or spread to other parts of the body. In epidemiological contexts, cancer prevalence refers to the proportion of a population found to have specific malignancies during a given period.
COPD	Chronic Obstructive Pulmonary Disease	A progressive lung disease characterized by persistent respiratory symptoms and airflow limitation due to airway and/or alveolar abnormalities usually caused by significant exposure to noxious particles or gases. COPD includes emphysema and chronic bronchitis, and is predominantly caused by smoking, air pollution, and occupational exposures.
CVD	Cardiovascular Disease	A group of disorders of the heart and blood vessels, including coronary heart disease, cerebrovascular disease, and other conditions affecting the cardiovascular system. In this study, CVD encompasses conditions that affect the heart's structure, function, and the circulatory system, which are leading causes of mortality globally.

Abbreviation	Term	Definition
DM	Diabetes Mellitus	A disease in which the body's ability to produce or respond to the hormone insulin is impaired, resulting in abnormal metabolism of carbohydrates and elevated levels of glucose in the blood and urine. In epidemiological studies, diabetes is typically categorized as Type 1, Type 2, or gestational, with Type 2 being most strongly associated with social determinants.
GAMM	Generalized Additive Mixed Model	A statistical model that extends generalized linear models to include nonlinear smoothing functions and random effects, used for analyzing complex data structures with non-linear relationships. GAMMs are particularly useful for environmental and public health data where relationships between variables often follow non-linear patterns.
GBDT	Gradient Boosting Decision Tree	An ensemble machine learning technique that combines multiple decision trees sequentially to correct errors made by previous models, commonly used for regression and classification problems. GBDT iteratively builds new models that predict the residuals or errors of prior models, then combines them to make a final prediction.
HTN	Hypertension	A condition in which the blood pressure in the arteries is persistently elevated above normal ranges (typically >130/80 mmHg), increasing the risk of heart disease, stroke, and other health problems. Hypertension is often called the "silent killer" due to its asymptomatic nature despite being a major risk factor for cardiovascular disease.
LR	Linear Regression	A statistical approach for modeling the relationship between a dependent variable (Y) and one or more independent variables (X) by fitting a linear equation $Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$ to observed data, where β values represent coefficients and ε represents error terms.

Abbreviation	Term	Definition
MAE	Mean Absolute Error	A metric used to evaluate the average magnitude of errors in a set of predictions, without considering their direction. Calculated as MAE = (1/n) Σ|y_i - ŷ_i|, where y_i are actual values and ŷ_i are predicted values. Unlike RMSE, MAE gives equal weight to all errors regardless of magnitude.
MAPE	Mean Absolute Percentage Error	A measure of prediction accuracy that expresses accuracy as a percentage of error, calculated as MAPE = (100/n) Σ|((y_i - ŷ_i)/y_i)|, where y_i are actual values and ŷ_i are predicted values. MAPE is scale-independent but can be distorted when actual values approach zero.
MLM	Multilevel Modeling	A statistical approach that analyzes hierarchical or nested data structures by accounting for variation at different levels, such as individual, group, or regional levels. MLM is particularly valuable in social determinants research where factors operate across individual, community, and societal levels.
MOPH	Ministry of Public Health	The governmental department in Thailand responsible for public health policies, healthcare services, and healthcare facilities management. MOPH serves as the primary source of health statistics and administrative data used in Thai public health research.
N/A	Medical Factors	Individual-level clinical and biological indicators used in healthcare settings to assess disease risk, progression, or outcomes, including blood pressure, cholesterol levels, blood glucose, BMI, family history, genetic markers, and other physiological measurements. These factors typically require direct clinical measurement or observation.
N/A	Non-Medical Factors	Factors outside the traditional clinical domain that influence health outcomes, including socioeconomic, environmental, behavioral, and cultural determinants that operate at both individual and population levels. In public health research, these factors are increasingly recognized as equally or more influential than medical factors in determining health outcomes.

Abbreviation	Term	Definition
N/A	Population-Level	Data, interventions, or analyses that address groups or communities rather than individuals, typically aggregated at geographic or demographic levels such as provinces, regions, or socioeconomic strata. Population-level approaches focus on patterns, distributions, and systemic factors rather than individual characteristics.
NCD	Noncommunicable Diseases	Medical conditions that are not infectious or transmissible from one person to another. These diseases tend to develop slowly over time and are often associated with long-term exposure to risk factors such as unhealthy diets, physical inactivity, tobacco use, and excessive alcohol consumption. NCDs include cardiovascular diseases, cancers, chronic respiratory diseases, and diabetes, which collectively account for over 70% of global deaths annually.
NSO	National Statistical Office	The principal government agency in Thailand responsible for collecting, processing, and disseminating statistical data and information. NSO conducts regular population surveys and censuses that provide critical socioeconomic data for public health research.
ODPHP	Office of Disease Prevention and Health Promotion	A U.S. federal office that provides leadership for disease prevention and health promotion programs and policies, whose SDH framework was adapted for this study. ODPHP developed the five-domain SDH framework (economic stability, education access and quality, healthcare access and quality, neighborhood and built environment, and social and community context) used to categorize features in this research.
pm2.5	Particulate Matter 2.5	Fine inhalable particles with diameters generally 2.5 micrometers and smaller, considered an important environmental health indicator and air pollution measure. PM2.5 particles can penetrate deep into the lungs and bloodstream, causing respiratory and cardiovascular diseases, making them a critical SDH factor in the built environment domain.

Abbreviation	Term	Definition
QALY	Quality Adjusted Life Year	A measure of health outcome that combines both the quantity and quality of life lived, calculated by multiplying life years by a utility value (between 0-1) representing health status. QALYs are used to evaluate the effectiveness and cost-effectiveness of medical interventions or treatments, with one QALY representing one year of perfect health.
RF	Random Forest	An ensemble learning method that operates by constructing multiple decision trees during training and outputting the average prediction of the individual trees for regression tasks. RF incorporates techniques like bagging and feature randomness to create uncorrelated forests of trees whose predictions are more accurate than those of individual trees.
RMSE	Root Mean Square Error	A standard way to measure the error of a model in predicting quantitative data, calculated as RMSE = $\sqrt{[(1/n)\ \Sigma(y\_i - \hat{y}\_i)^2]}$, where y_i are actual values and $\hat{y}\_i$ are predicted values. RMSE gives higher weight to larger errors due to the squaring operation, making it especially sensitive to outliers.
R^2	R Squared	A statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables, calculated as $R^2 = 1 -$ (Sum of Squared Residuals/Total Sum of Squares). R^2 ranges from 0 to 1, with higher values indicating better model fit, though it can be artificially inflated by adding predictors.
SDG	Sustainable Development Goals	The SDGs are a set of 17 global goals adopted by the United Nations General Assembly in 2015 as part of the 2030 Agenda for Sustainable Development. SDG 3 ("Ensure healthy lives and promote well-being for all at all ages") explicitly addresses NCDs and their social determinants, providing a global policy framework for the issues examined in this study.

Abbreviation	Term	Definition
SDH	Social Determinants of Health	The non-medical factors that influence health outcomes, including the conditions in which people are born, grow, work, live, and age, and the wider set of forces and systems shaping daily life. SDH encompasses factors such as income, education, employment, housing, access to healthcare, social support, and the physical environment, which collectively have greater impact on health outcomes than healthcare or individual behaviors.
SFFS	Sequential Forward Floating Selection	A feature selection algorithm that builds up a feature subset incrementally by including and excluding features to find the optimal combination for predictive modeling. SFFS performs bidirectional search, allowing previously selected features to be discarded if they become less relevant after adding new features, making it more flexible than simpler sequential approaches.
SVR	Support Vector Regression	A machine learning technique that applies the principles of Support Vector Machines to regression problems, using a non-linear mapping to transform the original training data into a higher dimension where it seeks to find an optimal hyperplane that maximizes the margin while tolerating error within specified thresholds (epsilon).
N/A	ST Depression	A finding on an electrocardiogram, wherein the trace in the ST segment is abnormally low below the baseline. ST depression often indicates myocardial ischemia (insufficient blood flow to the heart muscle) and is a significant diagnostic indicator for coronary artery disease when observed during stress testing.
N/A	ST Segment	The region between the end of ventricular depolarization (QRS complex) and beginning of ventricular repolarization (T wave) on the electrocardiogram. The ST segment represents the period when the ventricles are contracting but no electrical current is flowing, and deviations from the baseline are important indicators of cardiac pathology.

Abbreviation	Term	Definition
WHO	World Health Organization	A specialized agency of the United Nations responsible for international public health, setting norms and standards, and monitoring global health trends. WHO has established the Commission on Social Determinants of Health and developed frameworks guiding the understanding of how social factors shape health outcomes globally.
XGBoost	Extreme Gradient Boosting	An optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable, widely used for regression and classification problems. XGBoost implements machine learning algorithms under the Gradient Boosting framework with enhancements including regularization to prevent overfitting and parallel processing for improved computational efficiency.

BIODATA

Name-Surname: Mr. Peat Winch

Email: bhoothp@gmail.com