Solid Waste Generation Prediction Model Development: Case Study Saensuk Municipality, Chonburi

KITTIYA THIBUY

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR MASTER DEGREE OF SCIENCE
IN DATA SCIENCE
FACULTY OF INFORMATICS
BURAPHA UNIVERSITY
2022

มหาวิทยาลัยบูรพา
BURAPHA UNIVERSITY

การพัฒนาโมเดลสำหรับการพยากรณ์การสร้างขยะมูลฝอยของเทศบาลเมืองแสนสุข จังหวัดชลบุรี

กฤติยา ธิบาย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการข้อมูล
คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา
2565

Solid Waste Generation Prediction Model Development: Case Study Saensuk Municipality, Chonburi

KITTIYA THIBUY

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR MASTER DEGREE OF SCIENCE
IN DATA SCIENCE
FACULTY OF INFORMATICS
BURAPHA UNIVERSITY
2022

The Thesis of Kittiya Thibuy has been approved by the examining committee to be partial fulfillment of the requirements for the Master Degree of Science in Data Science of Burapha University

Advisory Committee

Principal advisor

.....

(Prajaks Jitngernmadan)

Examining Committee

..... Principal examiner
(Assistant Professor Rattana Wetprasit)

..... Member
(Assistant Professor Krisana Chinnasarn)

..... Member
(Prajaks Jitngernmadan)

..... Dean of the Faculty of Informatics
(Assistant Professor Krisana Chinnasarn)

.....

This Thesis has been approved by Graduate School Burapha University to be partial fulfillment of the requirements for the Master Degree of Science in Data Science of Burapha University

..... Dean of Graduate School
(Associate Professor Dr. Nujjaree Chaimongkol)

.....

63910158: MAJOR: DATA SCIENCE; M.Sc. (DATA SCIENCE)

KEYWORDS: Municipal Solid Waste, Prediction model, Multiple Linear Regression Model, Solid waste generation, Waste management

KITTIYA THIBUY : SOLID WASTE GENERATION PREDICTION MODEL DEVELOPMENT: CASE STUDY SAENSUK MUNICIPALITY, CHONBURI. ADVISORY COMMITTEE: PRAJAKS JITNGERNMADAN, Ph.D. 2022.

The increasing of solid waste generation in each region requires the development of strategies and methods, as well as the implementation of suitable management. If the prediction is possible, a more accurate waste management plan can be developed accordingly. In this research, an optimal prediction model for predicting solid waste generation in Saensuk Municipality, Thailand, is investigated and considered. Based on local constraints and existing factors, the three prediction algorithms are selected and studied. The Relatively Correlation coefficient of these three algorithms suggests that the Multiple Linear Regression (MLR) is the most suitable algorithm with R-squared = 0.74623. Other algorithms, which are Support Vector Regression (SVR) and Long Short Term Memory (LSTM), have the R-squared = 0.67469 and R-squared = 0.56242 respectively. For the reassurance process, the prediction's graph is used. It confirms that the MLR is eminently suitable for optimal solid waste generation prediction in Saensuk Municipality, Chon Buri.

ACKNOWLEDGEMENTS

We would like to thank Department of Public Health and Environment Saensuk Municipality, Civil Registration Work Saensuk Municipality, Ministry of Tourism and Sports, Office of the Registrar Burapha University, and Civil Registration Work Saensuk Municipality for providing useful data used in this research. Furthermore, we thank Digital Media and Interaction Research Laboratory (DMI) at Faculty of Informatics, Burapha University for the supports and suggestions.

Kittiya Thibuy

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

มหาวิทยาลัยบูรพา
BURAPHA UNIVERSITY

CHAPTER 1
INTRODUCTION

1.1 Problem Descriptions

Solid waste is one of the biggest problems our world has to find the optimal solution for managing it. One would see it as a useless waste, while the others can use it as a raw material. The question here is how to manage this solid waste to get the most out of it. In the global situation, the Allied Market Research reported in their Global Waste Management Market report that its value will reach 2,483.0 Billion US Dollars by 2030. The report includes Municipal Waste, Industrial Waste, and Hazardous Waste. In Thailand, the Pollution Control Department reported in 2020 that Thailand generated solid waste 25.37 Million Tons countrywide. Only 8.36 Million Tons of this solid waste was in the reuse or recycling process, whilst the 4.25 Million Tons of solid waste were not deposited and managed in the right way. In the city size, Chon Buri in this case, the 2,683.70 Tons of solid waste is generated daily. Only 9.28% of this solid waste was reused or recycled. The remains were buried in the deposal site or landed in the environment without pretreatment. This is also one of the enormous problems for the environment. Amongst the solid wastes generated nowadays, the municipal solid waste takes the second place of the number of the wastes after the industrial waste. This needs the professional management in terms of policy and information technology due to the environmental, social and economic challenges, especially in a municipality where the community lives next to each other.

1.2 Statement of the Problem

Understanding household waste generation behavior patterns is an important component of effective waste management. The current problem related to increasing waste generation in each area requires the search for strategies and methods and appropriate management. For waste management planning, it is essential that reliable waste generation data is available and that it is readily available. Both the amount of waste and the factors that influence the increase in waste, these data can be used in further analysis to create models for forecasting the upcoming generation of the waste.

The study area of this research is Saensuk District, Chonburi Province, which is an area with a lot of tourists coming in because it is an important tourist attraction of Chonburi. From visiting the area, it was found that there were still problems in waste management, which would come from both tourists and local people. During festivals or important days, there will be a lot of waste. If there is waste management or forecasting, it can help municipalities manage waste more efficiently.

1.3 Research Objectives

1. To optimize an optimal prediction model for forecasting solid waste generation in Saensuk Municipality, Chonburi Province.
2. To optimize the selected prediction model's features based on constraints and existing factors.

1.4 Research Questions

How to create the best prediction model for analyzing waste generation prediction in Saensuk municipality under existing factors?

Due to different factors exist in different cities, the most suitable prediction model for solid waste generation has to be found. These factors depend on e.g. size of the city, number of population, residents' behaviors and activities, etc. So, it is one of the most challenging task to find out what is the most suitable solid waste generation prediction model for a city. In this case is the Saensuk Municiple, Chon Buri province, Thailand.

1.5 Research Methodology

We collected relevant existing data to use in predicting the increase in solid waste in Saensuk municipality in five parts: The monthly amount of waste in Saensuk Municipality 2011-2020, Number of tourists in Saensuk Municipality 2009-2019, Average monthly household income 2002-2019, The population of Saensuk Municipality 2000 - 2020, Number of new students in each year in Burapha University, Saensuk Municipality 2009-2019. At the same time, We reviewed the literature between 2007 and 2021 related to the development of models for predicting the

occurrence of waste. We will analyze and compare various models that have been applied and select the best one to use in modeling the waste generation forecast of Saensuk Municipality to obtain a forecast. Artificial intelligence methods, artificial intelligence techniques, material flow analysis and statistical analysis, and time series analysis are some of the most commonly used forecasting tools in waste management. The emergence of future waste is the most accurate that will provide information to aid management decision-making and when management is able to make the right decisions and to enable effective solid waste management and will help reduce the environmental impact in a sustainable way

1.6 Threats to Validity

The data is used to develop the prediction model for solid waste generation in the municipality of Saensuk, Chonburi Province. This model is intended for usage in everyday situations. which isn't meant to be used in unusual situations like Disasters, wars or pandemic scenario, that impacts people's waste generating behavior.

1.7 Work phases and Timeline

Table 1 Timeline of the research

Tasks	Time Periods							
	Year1				Year 2			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
1. Problem Definition	←	—	→					
2. Literature Review	←	—	—	—	→			
3. Methodology and Solution Conception Framework Design and Testing		←	—	→				
4. Proposal Presentation and Approval					↔			
5. Institutional Review Board (IRB) Submission and Approval						↔		
6. Data Collection and Analysis							↔	
7. Test, Feedback, and Improvement							↔	

Tasks	Time Periods							
	Year1				Year 2			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
8. Academic International Conference Participation and Publication Presentation							↔	
9. Thesis Defense Examination								↔

CHAPTER 2
THEORIES AND RELATED WORKS

2.1 Basics of municipal solid waste generation prediction

The current problem related to increasing solid waste generation in the different areas requires sophisticated strategies and methods in order to manage it properly. We studied and analyzed the 10 most relevant research works qualitatively, spanning from the year 2014 to 2021. From analyzing of each paper, we found that the motivation for waste management problems varies from an environmental point of view to economic interest. In the works, the authors tried to predict municipal solid waste generation using different machine learning models and frameworks. On one hand, some of them attempted to compare the models' accuracy using different techniques. On the other hand, some of them used a specific machine-learning model to investigate its accuracy and precision. Furthermore, different factors or variables have been included in the model development process. These factors include GDP growth rate, total population, GDP per capita, mean household income per month, unemployment rate, crude death rate, labor force, tourists arrivals, number of households, urban life expectancy, average schooling, etc. Effective municipal solid waste management can be achieved by predicting the occurrence of solid waste precisely. However, this procedure is extremely difficult due to the uncertainty and unavailability of the data. Table 2 shows the relevant research works and the machine learning models used to predict solid waste generation.

2.2 State of the Art

In Sri Lanka, D.M.S.H. Dissanayaka et al.(Dissanayaka & Vasanthapriyan, 2019) developed a model for forecasting solid waste generation by employing principal component analysis and Pearson correlation to determine the link between factors impacting solid waste creation (Dissanayaka & Vasanthapriyan, 2019). The following are the seven factors that were considered in the study: population, average monthly family income, birth rate, and labor fatality rate, unemployment rate, per capita GDP, USD, GDP growth rate, and number of tourists. According to their findings, the total

population, birth rate, and GDP growth rate are the three factors that influence a municipal solid waste generation. In the model development section, they employed three forms of machine learning in the model development section: ANN, random forest, and regression analysis. Machine learning and linear and nonlinear models are examples of this. The ANN model may be used to forecast the most accurate municipal solid trash in Sri Lanka, according to them. The correlation coefficient is $R^2 = 0.6973$, $R^2 = 0.9608$, and $R^2 = 0.9923$ for linear regression, random forest, and ANN, respectively.

In a similar study, Sun, N., et al. mentioned that the ANN, the nonlinear model, gives high accurate result compare with regression analysis , which is a linear model (Sun & Chungpaibulpatana, 2017). The following variables are Total number of residents, native people aged 15 to 59 years, total people aged 15 to 59 years, number of households, income per household, and number of tourists are considered as the influential variables.

In Bogota, Colombia, Meza et al. examined artificial intelligence models to estimate solid waste generation. The following four variables were investigated: socioeconomic stratification, monthly solid waste generation, population, and the distribution of solid waste created by the urban region (Meza, Yepes, Rodrigo-Ilarri, & Cassiraga, 2019). They use a decision tree structure, a support vector machine, and a recurrent neural network model in this research. The best method for predicting solid waste generation was determined by comparing these three models. A decision tree is a tool for analyzing data. It is machine learning as a nonparametric algorithm that simulates data extraction constraints based on learning decision rules about the model's input behavior. Support vector machines are used as solid waste forecasting models. Even if there are issues with the data in the training phase, the benefit is that it allows for accurate data adjustment. The recurrent neural network model was used to investigate temporal connections among the same. According to the data presented in this paper, the appropriate predictor of future waste generation is SVM.

India's capital, New Delhi. (Soni, Roy, Verma, & Jain, 2019) They compared artificial intelligence models to see which one could predict waste generation in respective towns the best. There have been six different models used such as the artificial neural network (ANN), adaptive neuro-fuzzy inference system (ANFIS), and ANFIS and ANN coupled with genetic algorithm (GA) and discrete wavelet theory (DWT). To compare the models, the coefficient of determination (R^2), the root mean square error (RMSE), and index of agreement (WI) were calculated for each. Because it had the lowest RMSE (95.7) and the highest R2(0.87) and WI (0.864) values, the hybrid ANN–GA model was found to be the most accurate of the six models.
There are numerous studies on the application of the ANN model, which, according to research findings, is a model with good forecasting performance.

In Poland, (Elshaboury, Mohammed Abdelkader, Alfalah, & Al-Sakkaf, 2021) a team of researchers applied an optimized neural network to predict waste generation. The coefficient of efficiency (CE), Pearson correlation coefficient (R), Willmott's index of agreement (WI), root mean squared error (RMSE) and mean bias error (MBE) were used as a measurement for comparing the efficiency of ANN and ANN used with particle swarm optimization (PSO). In their research, five data were used as inputs for forecasting solid waste: population, employment to population ratio, revenue per capita, number of entities by type of business activity, and number of entities enlisted. in REGON per 10,000 population. Based on the research findings, it appears that the ANN–PSO model is a better and more cost-effective technique to design a waste management system than the traditional ANN model when comparing the values derived from various indicators.

In a similar study, Nguyen et al. (Nguyen et al., 2021) researchers in Vietnam compared six artificial intelligence models, including the linear model, support vector machine, cubist, random forest, and k-nearest neighbor, using various input inputs. According to their findings, the urban population, average monthly consumption expenditures, and total retail sales are all increasing. These three variables were determined to be the most essential in forecasting solid waste generation in Vietnam. MAE, RMSE, R2 are

indicators that can be used to analyze test results. The most efficient algorithms were reported to be random forest and k nearest neighbor.

Sun, N., et al. is mentioned that the ANN, the nonlinear model, gives high accurate result compare with regression analysis, which is a linear model. The following variables are Total number of residents, native people aged 15 to 59 year, total people aged 15 to 59 years, number of households, income per household, and number of tourists are considered as the influential variables (Sun & Chungpaibulpatana, 2017).

2.3 Tools

The tool used for data manipulation and machine learning in this work is Google Collaboratory (Colab, 2022), which is used to write Python code to create a Machine Learning Model. Google Collaboratory is popular and suitable for tasks related to data management and report generation. The library can be imported, and the results can be seen immediately. The libraries that are important in data estimation are:

1. Numpy that is used to calculate various mathematical functions. (NumPy, 2022),
2. Pandas that are used to manage data (NumFOCUS, 2022). It can help to complete data preparation before it can be used in calculations as well as being able to read files and write files in various formats,
3. Matplotlib that is used to create basic graphs with a wide variety of graphs (Matplotlib, 2022). If you want to make the graphs more beautiful, you can use the library called,
4. Seaborn (Waskom, 2022), which is a library developed from Matplotlib, and
5. Scikit-learn (sklearn) (Scikit-learn, 2022a) that is used in machine learning because it can handle more data such as classification, segmentation, regression analysis, support vector machine, decision tree and k-mean.

2.4 Algorithms

1. PC: Pearson Correlation

   The purpose of the PC (Wikipedia, 2022b) is to define the relationship between two variables on an interval or ratio scale. The value is referred to. In most cases, the "correlation coefficient" is between -1.00 and 1.00.

- If a negative value is present, it indicates that the two variables are related in the other manner.
- If the value is positive, the two variables have the same direction of relationship.
- The two variables are unrelated if the value is 0.

2. PCA: Principal Component Analysis

   Principal Component Analysis is a multivariate data analysis technique that does not divide variables into dependent and independent variables(Wikipedia, 2022c). It simply concerns about discovering the relationship between those variables and creating a new variable made up of the original variable's variation or variance. As a result, this kind of analysis is frequently used to reduce the size of the variables' matrix or to correlate the data. Several papers has used the PCA technique to artificial neural network (ANN) procedures with the goal of reducing analysis time and optimizing the process.

3. LRM: Linear Regression Model

   Linear Model include to Linear Regression is a technique from statistics and Machine Learning algorithm (MathWorks, 2022). A regression function with a defined set of input values (x) and forecasted output for that set of input values(y) is called Linear Regression. Although Linear Regression and Machine Learning are simple to understand and interpret, they do have significant restrictions on independent and normal distribution resulting in a substantial bias in the output. Because of the linear nature of this method, it is unsuitable for modeling highly non-linear data. Ordinary least squares linear regression's goal is to find the plane with the lowest sum-of-squared errors between observed and forecasted responses. Because of the simple underlying method and simplicity of interpretation of the findings, LM is an appealing model for practical applications; however, correlation among the predictors might result in LM with large errors.

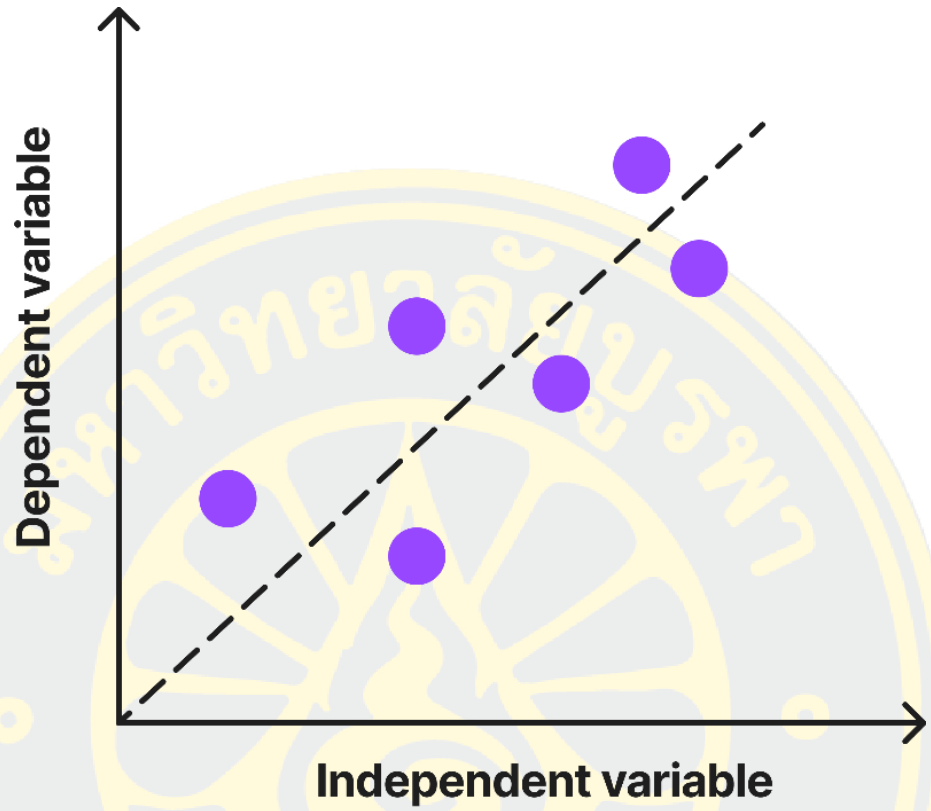


Figure 1 Linear Regression Model

4. ANN: Artificial Neural Networks

An artificial neural network (ANN) is a model of neural networks in the human brain to give computers the ability to learn and recognize patterns (Wikipedia, 2022a). The human brain has many microprocessors and is connected by neural networks that allow it to learn and think quickly, but computers are not as complex as the brains of the human brain. Humans are only responsible for running programs according to human instructions. In order for computers to learn, it is difficult to simulate the way people learn to computers to form an artificial neural network. ANN models are very useful to Curve fitting, classification, clustering, and dynamic time series forecasting. Because of its high fault tolerance capability and suitability for portraying complex relationships between variables in multivariate systems, ANN has been used in numerous waste management research.

There are many types of ANN architecture.

- Classification (Supervised learning) such as feedforward networks (MLP), backpropagation, and radial basis function networks
- Clustering (Unsupervised Learning), such as adaptive Resonance Theory (ART) and self-organizing map (SOM).
- Association (Unsupervised Learning), such as Hopfield networks
- Architecture that is commonly applied is Feedforward, a multi-layered perceptron with a propagation learning algorithm that is used for classification or regression.

A Supervised Learning Process

a. Compute temporary outputs
b. compare outputs with desires targets
c. adjust the weights and repeat the process

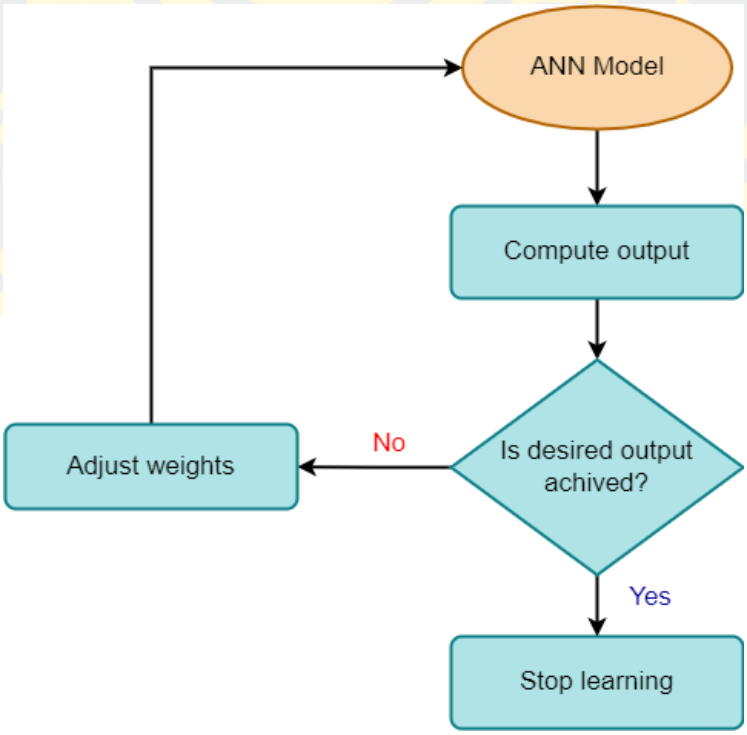


Figure 2 A Supervised Learning Process

5. RFM: Random Forest Model

   Random Forests are an improvement over bagged decision trees. A problem with decision trees like CART (Wikipedia, 2022d). They choose which variable to split on using a greedy algorithm that minimizes error. As such, the decision trees can have a lot of structural similarities and in turn result in high correlation in their predictions. Combining predictions from multiple models in ensembles works better if the predictions from the sub-models are uncorrelated or at best weakly correlated. Random forest changes the algorithm for the way that the sub-trees are learned so that he resulting predictions from all of the subtrees have less correlation. It is a simple tweak. In CART, when selecting a split point, the learning algorithm is allowed to look through all variables and all variable values in order to select the most optimal split-point. The random forest algorithm changes this procedure so that the learning algorithm is limited to a random sample of features of which to search. The number of features that can be searched at each split point (m) must be specified as a parameter to the algorithm.

6. SVM: Support vector machine

   A SVM finds the hyper-plane that best separates the two classes by plotting each data item as a point in n-dimensional space (n being the number of features) (Scikit-learn, 2022b). SVM is particularly good at solving classification problems and is more commonly used to forecast discrete outputs It's been utilized to handle regression issues including MSW forecast, waste sorting, and energy recovery, among others This approach can be used for numeric prediction and classification. In the literature, a combination of SVM and PCA has been used to forecast the weekly creation of solid garbage in Mashhad, Iran.

2.4 The Published Prediction Models

The effective municipal solid waste management can be achieved by predicting the occurrence of solid waste precisely. However, this procedure is extremely difficult due

to the uncertainty and unavailability of the data. Table 2 shows the relevant research works and the machine learning models used to predict the solid waste generation.

Table 2 Relevant research works and the machine learning models used to predict the solid waste generation

Work	Year	Method	Duration	Variable	Result model	Area
1	2019	• PC • PCA • LRA • ANN • RFM	yearly	• Crude Birth Rate • GDP • Total Population • GDP per capita USD • Mean household income per month • Unemployment Rate • Crude Death Rate • Labor Force • Tourists Arrivals	ANN	Municipality
2	2017	• PC • PCA • LRA • ANN	yearly	• Total MSW • Total number of residents • Native residents • Native people aged 15-59 years • Total people aged 15-59 years • Number of households • Income per household • Number of tourists	ANN	Municipality
3	2016	• LRA • TSA	year	• The number of inhabitants • Population aged 15to 59 years • Urban life expectancy	LRA	Municipality

Work	Year	Method	Duration	Variable	Result model	Area
				• Amounts of MSW generate 2000- 2014 • Infant mortality rate and life expectancy at birth • Household size • GDP • The labor force in agriculture		
4	2019	• LSVM • SVM • DT	day	• Distribution by collection zone in the city • Socio-economic stratification • Population • Quantity of solid waste generated in a determined period of time	SVM	City
5	2021	• LM • SVM • Cubist • RFM • k-NN	day	• Total solid waste collected per day in the province • The monthly average income per capita • Average monthly consumption expenditure per capita • Number of people per unit area • Average size (area) occupied in the home per person • Total retail sales of consumer goods per year • Urban population in towns and cities per province • Total hospital beds • Total residential land area	SVM	Country

Work	Year	Method	Duration	Variable	Result model	Area
6	2017	• GRNN • SB-GRNN	year	• GDP • Domestic material consumption • Urban population • Population • Average household size • Industry, value added • Tourism expenditure in the country • Population by age group 20-65 • Unemployment rates • Alcohol consumption • Household final consumption expenditure • CO2 emissions from residential buildings and com. and public services	SB-GRNN	Country
7	2019	• ANN • DWT-ANN • GA-ANN • Pure ANFIS • DWT-ANFIS • GA-ANFIS	year	• MSW generation	GA-ANN	City
8	2021	• ANN	year	• Population	**ANN-PSO**	Municipality

Work	Year	Method	Duration	Variable	Result model	Area
		• **ANN-PSO**		• Employment to population ratio • Revenue per capita • Number of entities by type of business activity • Number of entities enlisted in REGON per 10,000 population.		

From the analysis of these 8 documents, we found that in total 17 machine-learning models were used for data preparation, analysis, and solid waste generation forecasting. The figure 1 shows the frequency of the used machine learning models.

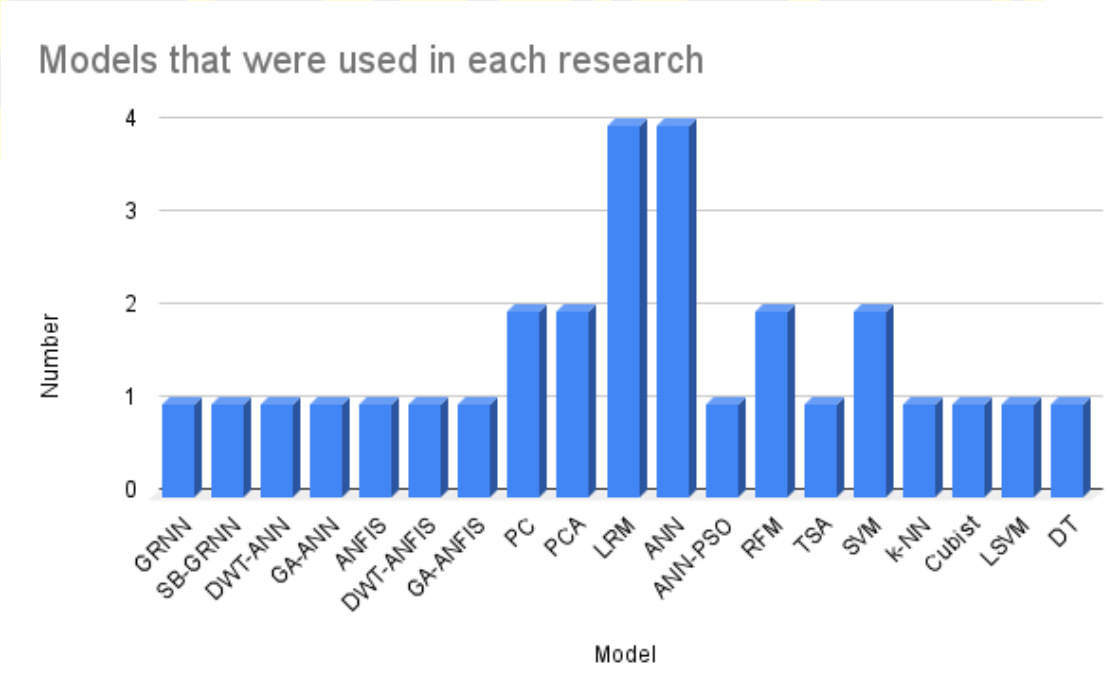


Figure 3 The frequency of each algorithm used in the reviewed works.

Where as:

1. GRNN: General Regression Neural Network
2. SB-GRNN: Structural Breaks General Regression Neural Network
3. DWT-ANN: Discrete Wavelet Theory–Artificial Neural Network
4. GA-ANN: Genetic Algorithm–Artificial Neural Network
5. GA-ANFIS: Genetic Algorithm–Adaptive Neuro-Fuzzy Inference System
6. DWT-ANFIS: Discrete Wavelet Theory–Adaptive Neuro Fuzzy Inference System
7. ANFIS: Adaptive Neuro-Fuzzy Inference System
8. PC: Pearson Correlation
9. PCA: Principal Component Analysis
10. LRM: Linear Regression Model
11. ANN: Artificial Neural Networks
12. ANN-PSO: Artificial Neural Networks–Particle Swarm Optimization
13. RFM: Random Forest Models
14. TSA: Time Series Analysis
15. SVM: Support Vector Machine
16. k-NN: k-Nearest Neighbors
17. Cubist: Cubist Regresion Models
18. LSVM: Linear Support Vector Machine
19. DT: Decision Tree

The analysis of these research works shows that 17 different models were used in terms of solid waste generation prediction, whereas the most used model is the one with the Artificial Neural Network (ANN) method. This model was applied to 4 documents, which worked quite well for the data series used, both yearly and daily ones. Furthermore, this model could also be used in the city and municipality area. The Support Vector Machine (SVM) method was applied to 2 documents, which worked well for day-to-day data series. This model can also be used in the city and country area. The Pearson Correlation (PC), The Principal Component Analysis (PCA), Time Series Analysis (TSA) method was applied to 2 documents, which worked well only for data series that were recorded with yearly time period. This model can also be used in the municipal area. The Random Forest Model (RFM) method was applied to 2 documents, which worked well for yearly and daily data sources. This model can also

be used in country and municipal areas. The Linear Regression Model (LRM) model method was applied to 4 documents, which worked well for yearly and daily data sources. This model can also be used in the country and municipal areas. Furthermore, 2 methods were used in the country areas with daily data collection, which are the Cubist and the k-Nearest Neighbors (k-NN). In addition, 2 methods were used in the country areas with yearly data collection, namely the General Regression Neural Network (GRNN), the Structural Breaks General Regression Neural Network (SB-GRNN), The 2 methods were used in the city areas with daily data collection, namely Decision Tree (DT) and Linear Support Vector Machine (LSVM) Finally, 4 methods were used in the city areas with yearly data collection that are the Discrete Wavelet Theory–Artificial Neural Network (DWT-ANN), the Discrete Wavelet Theory–Adaptive Neuro-Fuzzy Inference System (DWT-ANFIS), the Genetic Algorithm–Artificial Neural Network (GA-ANN), and the Genetic Algorithm–Adaptive Neuro-Fuzzy Inference System (GA-ANFIS).

2.5 The area size and the suitable models

The size of the research area plays an important role, too. The bigger area has more varieties in population and activities. This affects the way the solid waste will be generated. Especially, in the urban area where the people usually consume more products and stay longer in the night. Another factor affected the prediction model selection is the interval of data collection. Some prediction models are suitable for precise data collection while some can be applied to approximate data series. Table 3 shows the area size and the suitable models.

Table 3 The area size and the suitable models

Area Size	Type	Yearly	Daily
Large	country	GRNN	LRM
		SB-GRNN	RFM
			SVM
			k-NN
			CUBIST

Area Size	Type	Yearly	Daily
Middle	city	ANN	LSVM
		DWT-ANN	***SVM***
		GA-ANN	DT
		ANFIS	
		DWT-ANFIS	
		GA-ANFIS	
Small	municipal	PC	
		PCA	
		LRM	
		ANN	
		ANN-PSO	
		RFM	
		TSA	

2.6 The Factors/Variables

The factors or variables are one of the most important components affecting the accuracy of the solid waste generation prediction. The people and their activities can lead to the solid waste generation, i.e., an area with ecoconscious living people would generate less solid waste than in an area with average people. In our analysis, we found that almost of the research works included the population or number of residents as a factor in the calculation. Another interesting factor is the number of the tourist arrivals in that area. The tourists may generate more solid waste because they may use disposable products more than the residents may. Other factors like GDP or GDP per capita and occupation may reflect how people live their daily life. The more GDP per capita, the less is the solid waste generation. Furthermore, the level of education and the living area affect the way people generate the solid waste, too. The people with lower education level or living in a urban area tend to generate more solid waste. These

factors have to be included in the framework building for municipal solid waste generation prediction.

2.7 Gap Identification

The presented models above is utilized with a variety of factors and different area sizes. Each of these three portions is unique, which means that each area is associated with respective factors that have a specific impact on that area. As a result of the different area sizes and circumstances, the prediction model developed in other places cannot be used in the Saensuk Municipality area. As a result, the prediction model must be developed in the Saensuk area using factor data from Saensuk Municipality. The information about Saensuk's elements is specific, such as the presence of tourists and the number of passive populations that come from universities and personnel. For this reason, we conducted this study in order to design the best appropriate model for prediction waste generation in Saensuk Municipality, Chonburi Province, based on constraints and determinants.

CHAPTER 3
METHODOLOGY AND ANALYSIS

In Chapter 3, an experimental design was developed. The different areas with different conditions must use different approaches with a proper prediction model.

There are 3 most important components concerning municipal solid waste generation prediction, which are

1. The most suitable prediction model.
2. The size of the study area.
3. The factors/variables applied.

In the case of small area with yearly data interval collection, the Artificial Neuron Networks (ANN) model and The Linear Regression Model (LRM) mode are the most suitable for solid waste generation prediction, under the conditions that the suitable variables are selected. These variables could be the number of residents, number of tourists, GDP, GDP per capita, educational background, and living area such as urban or land area. The variable “number of tourists” plays an important role in our future work. We will use this state of the art as a basic for our municipal solid waste generation prediction approach. The research area will be the Saensuk Municipality, which is next to the Bangsaen Beach and welcomes many tourists, especially on weekends. This prediction model should help the municipal office planning how to manage the solid waste that would be generated in the future. Which in the overview of working in the experimental design as follows.

3.1 Overview of the Experiment

The concept of this experiment is based on the training and testing for creating a prediction model. The overview of experiment method is shown in the figure 4 below.

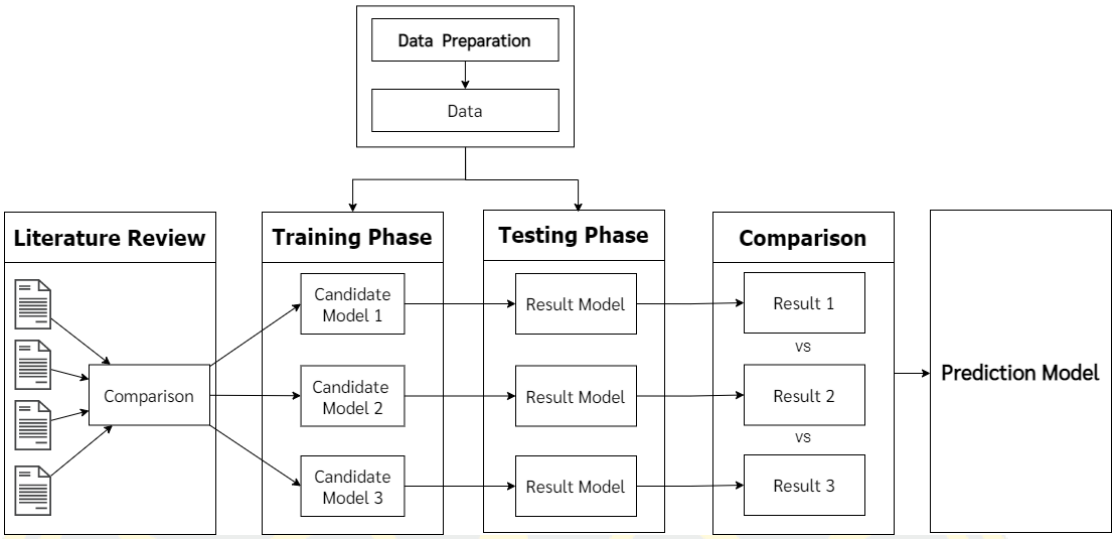


Figure 4 Overview of the Experiment

Figure 4 depicts the process of the workflow which is we can be described as follows:

- Firstly, From the study of the documents, we searched for the documents using the keyword "Solid Waste Generation Model" and then separated the documents according to the size of the area and the frequency of the data.

  There are 3 types of area size which are:

  1. Country
  2. City
  3. Municipality

  There are 2 types of data frequency which are:

  1. Daily
  2. Yearly

  When grouped, the best model for forecasting municipal solid waste with yearly inputs for various factors was Artificial Neuron Networks model and the Linear Regression Model.
- Secondly, Data preparation means any process that needs to be done with the raw data obtained. Then modify the data to a suitable format to be loaded into the model or analyzed further.
- Thirdly, Take the available data and feed it into the model, which will build a prediction model and then return the model's result in training phase.

- Finally, The results of each model were compared to obtain the best model for predicting the generation of solid waste in Saensuk Municipality, Chonburi Province.

3.2 Study Area

Areas of Saensuk Municipality, Chonburi Province The city council of Saensuk Responsible for taking care of a total area of 20.286 square kilometers (S. M. Office, 2022a). In this area, most of the land use was the top three being road and rail areas, the first was used for residential areas, and the third was for educational institutions. The population in this area has a total of 46,131 people and 37,588 houses, there are 12,312 households, the average household size is 4 people per household.

The distinctive feature of Saensuk Municipality is that it is an important tourist attraction of Chonburi Province. Therefore, there are a large number of tourism-related businesses such as hotels, shops, restaurants, which will bring a large flow of people into the area. There is also a university Which is the center of education in the eastern region of Thailand, so there is a large number of people who come to live to work and study. As a result, there is a large number of unidentified populations in the local register, and the municipality estimates that there are over 10000 hidden populations in the Saensuk municipality.

3.3 An Approach

Phase 1: Data Collection

The collection of raw data from Seansuk Municipality is featured with various significant factors. These factors are determined from our studies in the section II mentioned above. These selected 5 factors are similar to the case studies, as well as the size of the city, the population, the behavior and activities of the residents. The data was collected from sources existing in different departments of Saensuk Municipality, see Table 4.

Table 4 Factors Data

Factor	Unit	Frequency	Source of data
Solid waste	Ton/month	Monthly	Department of Public Health and Environment Saensuk Municipality
Average monthly household income	Baht/Year	Yearly	Civil Registration Work Saensuk Municipality. (N. S. Office, 2022)
Tourists	Person/Year	Yearly	Ministry of Tourism and Sports (Sports, 2022)
Population	Person/Year	Yearly	Civil Registration Work Saensuk Municipality (S. M. Office, 2022b)
Student Enrollment	Person/Year	Yearly	Office of the Registrar Burapha University (Registrar, 2022)

These 5 hyperlocal factors affect solid waste generation in Saensuk. The data is from 2009 – 2019:

1. Monthly solid waste data We have requested information from Department of Public Health and Environment Saensuk Municipality.

ข้อมูลปริมาณขยะที่นำมากำจัดที่ศูนย์กำจัดขยะมูลฝอยของเทศบาลเมืองแสนสุข ปี งบประมาณ 2563

เดือน	ปีงบประมาณ พ.ศ. 2563					
	ทม.แสนสุข		อบต.บางพระ		รวมทั้งหมด	
	(ตัน/เดือน)	(ตัน/วัน)	(ตัน/เดือน)	(ตัน/วัน)	(ตัน/เดือน)	(ตัน/วัน)
Oct-62	3,463.110	111.713	1,291.870	41.673	4,754.980	153.386
Nov-62	3,136.870	104.562	1,142.630	38.088	4,279.500	142.650
Dec-62	3,328.370	107.367	758.450	24.466	4,086.820	131.833
Jan-63	3,385.770	109.218	845.060	27.260	4,230.830	136.478
Feb-63	3,056.490	105.396	635.600	21.917	3,692.090	127.313
Mar-63	3,127.750	100.895	771.980	24.903	3,899.730	125.798
Apr-63	2,621.280	87.376	770.970	25.699	3,392.250	113.075
May-63	2,922.550	94.276	878.760	28.347	3,801.310	122.623
Jun-63	3,223.690	107.456	872.940	29.098	4,096.630	136.554
Jul-63	3,437.380	110.883	846.660	27.312	4,284.040	138.195
Aug-63	3,629.740	117.088	812.190	26.200	4,441.930	143.288
Sep-63	3,326.310	110.877	803.790	26.793	4,130.100	137.670
รวม	**38,659.310**	**1,267.109**	**10,430.900**	**341.755**	**49,090.210**	**1,608.864**
เฉลี่ย (ตัน/เดือน)	**3,221.609**	**105.592**	**869.242**	**28.480**	**4,090.851**	**134.072**
สรุปปริมาณขยะทั้ง 2 หน่วยงาน	**(ตัน)**			**49,090.210**		
	(ตัน/เดือน)			**4,090.851**		

Figure 5 Example of solid waste data

Because we must submit the agency documentation asking for information. The team has organized the data into Excel files that are easily usable. Due to the fact that the acquired information pertains to regions other than our study area, the data must be cleansed by removing just the Saensuk related information.

2. Yearly data of the average monthly household income of the population in Saensuk municipality received information from Civil Registration Work Saensuk Municipality.

ค่าใช้จ่ายเฉลี่ยต่อเดือนของครัวเรือน เป็นรายภาค และจังหวัด พ.ศ. 2555 - 2564

หน่วย: บาท

ภาค	จังหวัด	2555	2556	2557	2558	2559	2560	2561	2562	2563	2564
ทั่วราชอาณาจักร	ทั่วราชอาณาจักร	18,766.00	19,061.00	20,892.00	21,157.00	21,144.00	21,436.50	21,346.00	20,742.12	21,329.00	21,616.00
กรุงเทพมหานคร และ 3 จังหวัด	กรุงเทพมหานคร และ 3 จังหวัด	31,971.00	32,425.00	31,606.00	30,882.00	32,091.00	33,126.02	33,408.00	30,778.10	31,142.00	31,382.00
	กรุงเทพมหานคร	33,956.98	35,023.70	34,425.64	33,085.70	35,101.40	35,350.70	34,127.44	31,753.04	32,052.03	31,866.68
	สมุทรปราการ	25,860.88	26,192.90	22,747.21	22,331.80	24,353.97	24,354.72	23,231.71	21,423.43	23,850.82	27,484.76
	นนทบุรี	28,731.23	26,946.60	30,812.06	31,381.00	28,828.37	33,313.04	33,808.98	32,189.09	33,042.31	33,995.57
	ปทุมธานี	30,668.81	29,514.00	30,197.07	29,770.00	31,271.04	33,604.46	43,300.51	37,086.11	33,823.84	31,639.92
ภาคกลาง	ภาคกลาง	19,762.00	19,728.00	21,144.00	21,055.00	20,493.00	21,119.75	21,168.00	20,644.55	21,771.00	22,332.00
	พระนครศรีอยุธยา	25,215.89	20,493.70	20,409.63	22,218.10	23,095.32	23,780.19	22,790.08	24,439.76	23,391.26	25,326.92
	อ่างทอง	21,273.24	21,182.50	19,633.96	17,573.60	20,371.91	17,162.48	17,294.67	17,010.98	17,726.59	17,020.79
	ลพบุรี	17,356.96	15,944.80	14,910.71	17,968.90	15,875.50	16,012.15	17,016.83	16,829.71	19,180.09	21,324.54
	สิงห์บุรี	23,474.62	22,118.20	19,631.55	22,136.80	19,381.46	20,262.78	19,773.04	19,884.51	20,303.62	20,146.07
	ชัยนาท	15,535.57	17,766.60	17,090.91	17,163.10	17,495.47	16,162.12	17,187.28	15,762.99	19,835.74	18,292.27
	สระบุรี	21,412.68	22,765.00	22,811.16	23,017.10	23,964.34	26,635.10	27,581.38	26,007.12	26,045.40	26,503.14
	ชลบุรี	**25,499.01**	**24,934.20**	**25,704.10**	**24,182.00**	**24,257.06**	**24,572.50**	**25,322.87**	**25,683.70**	**24,878.44**	**28,001.46**
	ระยอง	21,023.90	21,872.50	23,303.13	24,433.50	21,024.65	22,698.79	19,409.81	20,806.85	21,451.36	22,365.56
	จันทบุรี	19,594.74	17,597.30	20,649.74	23,350.80	22,790.42	20,619.92	20,922.15	19,812.57	23,300.20	22,347.34
	ตราด	15,662.72	16,706.30	18,126.65	18,989.00	18,913.56	20,404.69	20,198.98	18,883.88	19,563.16	19,796.09
	ฉะเชิงเทรา	23,079.51	26,070.70	23,342.12	21,782.60	21,674.11	21,437.44	19,070.81	17,035.88	18,791.18	18,968.49
	ปราจีนบุรี	19,733.75	18,314.70	20,789.55	20,994.50	18,156.85	19,268.28	20,782.66	21,677.72	22,470.73	23,318.79
	นครนายก	16,459.52	17,696.90	17,482.59	18,153.50	17,877.62	18,601.09	19,152.88	19,717.14	20,609.12	21,775.45
	สระแก้ว	19,531.06	18,571.10	20,227.36	20,576.70	18,413.18	17,609.53	17,543.55	15,827.78	16,939.41	17,347.26

Figure 6 Example of average monthly household income data

This information is accessible online. The Excel file is saved for download and use convenience. Due to the fact that the acquired information pertains to regions other than our study area, the data must be cleansed by removing just the Saensuk related information.

3. The total population in the municipality of Saensuk on an annual basis received information from civil registration work Saensuk Municipality.

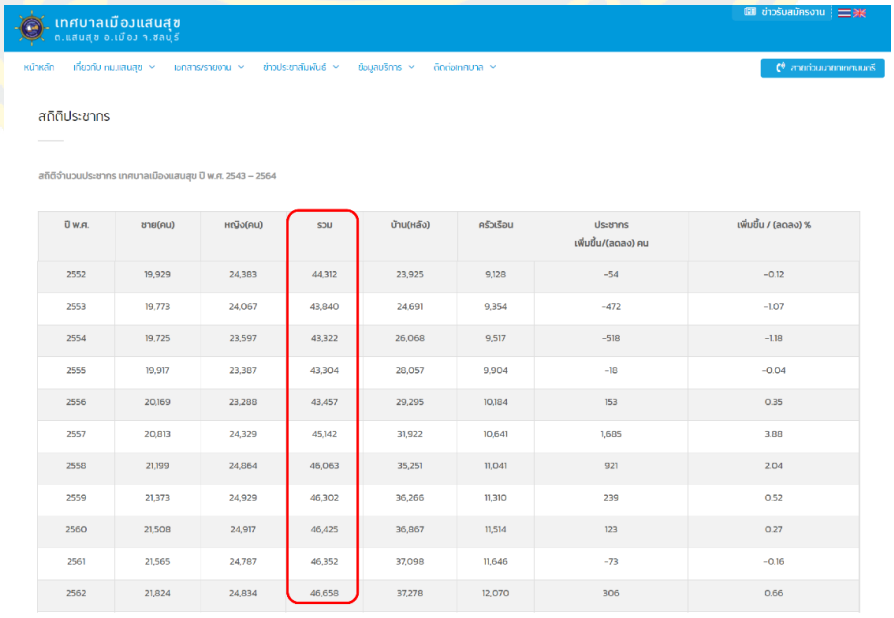
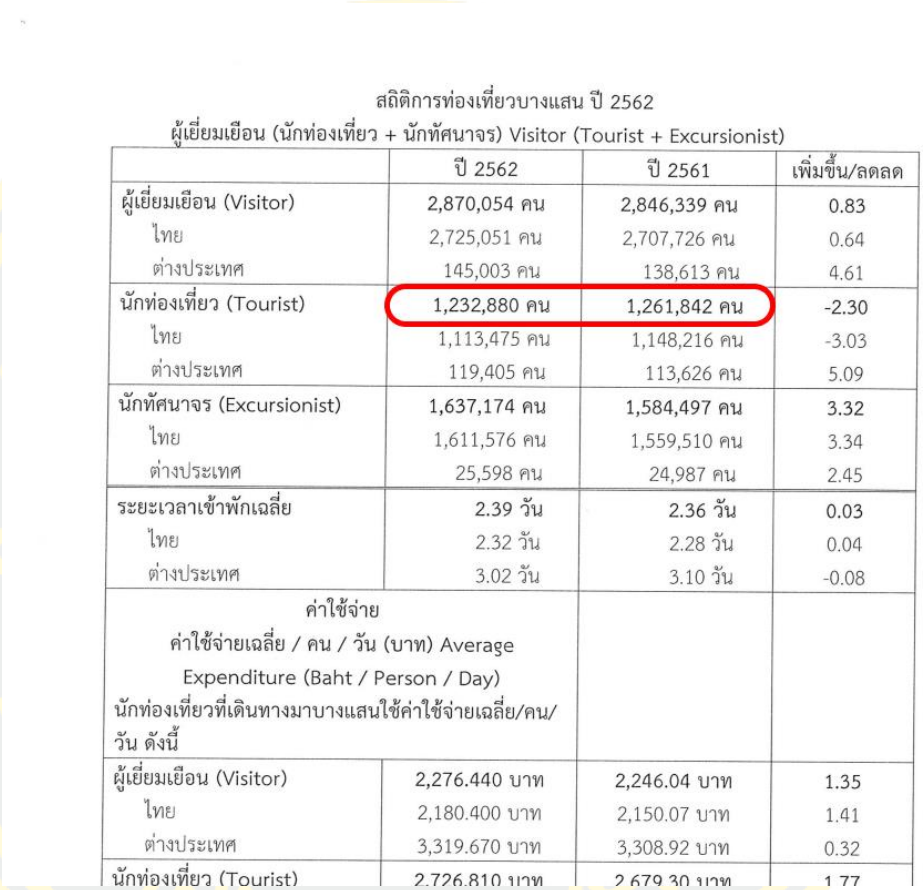


Figure 7 Example of total population data

This data is accessible online. Therefore, we only select information that is usable and has the same properties as other factors.

4. The annual total of tourists in Saensuk municipality was obtained from Ministry of Tourism and Sports.



สถิติการท่องเที่ยวบางแสน ปี 2562
ผู้เยี่ยมเยือน (นักท่องเที่ยว + นักทัศนาจร) Visitor (Tourist + Excursionist)

	ปี 2562	ปี 2561	เพิ่มขึ้น/ลดลด
ผู้เยี่ยมเยือน (Visitor)	2,870,054 คน	2,846,339 คน	0.83
ไทย	2,725,051 คน	2,707,726 คน	0.64
ต่างประเทศ	145,003 คน	138,613 คน	4.61
นักท่องเที่ยว (Tourist)	1,232,880 คน	1,261,842 คน	-2.30
ไทย	1,113,475 คน	1,148,216 คน	-3.03
ต่างประเทศ	119,405 คน	113,626 คน	5.09
นักทัศนาจร (Excursionist)	1,637,174 คน	1,584,497 คน	3.32
ไทย	1,611,576 คน	1,559,510 คน	3.34
ต่างประเทศ	25,598 คน	24,987 คน	2.45
ระยะเวลาเข้าพักเฉลี่ย	2.39 วัน	2.36 วัน	0.03
ไทย	2.32 วัน	2.28 วัน	0.04
ต่างประเทศ	3.02 วัน	3.10 วัน	-0.08
ค่าใช้จ่าย ค่าใช้จ่ายเฉลี่ย / คน / วัน (บาท) Average Expenditure (Baht / Person / Day) นักท่องเที่ยวที่เดินทางมาบางแสนใช้ค่าใช้จ่ายเฉลี่ย/คน/วัน ดังนี้			
ผู้เยี่ยมเยือน (Visitor)	2,276.440 บาท	2,246.04 บาท	1.35
ไทย	2,180.400 บาท	2,150.07 บาท	1.41
ต่างประเทศ	3,319.670 บาท	3,308.92 บาท	0.32
นักท่องเที่ยว (Tourist)	2,726.810 บาท	2,679.30 บาท	1.77

Figure 8 Example of total of tourists data

This data is accessible online. But the information written down on paper must be saved by hand to an Excel file. The image from paper copy data is a collection of data that can be hard to use.

5. Number of student enrollment Burapha University annually received information from Student registration and statistics Burapha University.

จำนวนผู้มีสิทธิ์เข้ารับพระราชทานปริญญาบัตร
ประจำปีการศึกษา 2561-2562
วันที่ 19-21 มกราคม พ.ศ. 2565

คณะ/วิทยาลัย	ปริญญาตรี			ปริญญาโท			ปริญญาเอก			รวมทั้งสิ้น
	ปกติ	พิเศษ	รวม	ปกติ	พิเศษ	รวม	ปกติ	พิเศษ	รวม	
คณะการจัดการและการท่องเที่ยว	586	925	1,511	-	89	89	-	-	0	1,600
คณะการแพทย์แผนไทยอภัยภูเบศร	150	-	150	-	-	0	-	-	0	150
คณะดนตรีและการแสดง	173	-	173	-	-	0	-	-	0	173
คณะเทคโนโลยีการเกษตร	49	-	49	-	-	0	-	-	0	49
คณะเทคโนโลยีทางทะเล	173	1	174	-	-	0	-	-	0	174
คณะพยาบาลศาสตร์	273	-	273	38	54	92	11	-	11	376
คณะแพทยศาสตร์	86	-	86	-	-	0	-	-	0	86
คณะภูมิสารสนเทศศาสตร์	241	212	453	6	-	6	7	-	7	466
คณะเภสัชศาสตร์	183	-	183	-	-	0	-	-	0	183
คณะมนุษยศาสตร์และสังคมศาสตร์	1,104	794	1,898	3	37	40	6	-	6	1,944
คณะรัฐศาสตร์และนิติศาสตร์	616	770	1,386	13	87	100	-	29	29	1,515
คณะโลจิสติกส์	392	316	708	2	85	87	7	-	7	802
คณะวิทยาการสารสนเทศ	690	41	731	8	-	8	1	-	1	740
คณะวิทยาศาสตร์	517	49	566	45	36	81	10	-	10	657
คณะวิทยาศาสตร์การกีฬา	415	-	415	7	-	7	15	-	15	437
คณะวิทยาศาสตร์และศิลปศาสตร์	443	-	443	-	-	0	-	-	0	443
คณะวิทยาศาสตร์และสังคมศาสตร์	360	28	388	-	-	0	-	-	0	388
คณะวิศวกรรมศาสตร์	675	156	831	30	32	62	3	-	3	896
คณะศิลปกรรมศาสตร์	252	73	325	-	18	18	48	-	48	391
คณะศึกษาศาสตร์	443	230	673	108	234	342	144	30	174	1,189
คณะสหเวชศาสตร์	257	-	257	3	-	3	1	-	1	261
คณะสาธารณสุขศาสตร์	295	29	324	3	27	30	2	-	2	356
คณะอัญมณี	163	1	164	-	-	0	-	-	0	164
โครงการจัดตั้งคณะพาณิชยศาสตร์และบริหารธุรกิจ	418	-	418	-	-	0	-	-	0	418
วิทยาลัยการบริหารรัฐกิจ	-	-	0	2	86	88	70	-	70	158
วิทยาลัยนานาชาติ	467	-	467	-	-	0	-	-	0	467
วิทยาลัยพาณิชยศาสตร์	-	-	0	12	121	133	1	60	61	194
วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา	-	-	0	24	-	24	73	-	73	97
ศูนย์ฝึกพาณิชยนาวี (สถาบันสมทบ)	165	-	165	-	-	0	-	-	0	165
รวมทั้งสิ้น	9,586	3,625	13,211	304	906	1,210	399	119	518	14,939

Figure 9 Example of student enrollment data

This information is accessible online. The Excel file is saved for download and use convenience. Therefore, we only select information that is usable and has the same properties as other factors.

Phase 2: Data Preparation

The data cleansing is the process of validating, correcting, or removing invalid entries from a dataset because the data comes from a variety of sources. Once the data has been correctly processed, it will be in the same format and ready to be used.

1. Data Cleansing

   Data cleansing is a process of validating, correcting, or deleting invalid entries from a dataset, table, or database, which serves as the database's cornerstone. Incompleteness, errors, irrelevance to other data, and so on are all examples of this. As a result, many experts consider data cleanup to be the most crucial part of data quality management. Because the data comes from a variety of sources. We need to gather all of the data in one place. Then look for any missing data. Once the data has been correctly processed, it will be in the same format and ready to be used.

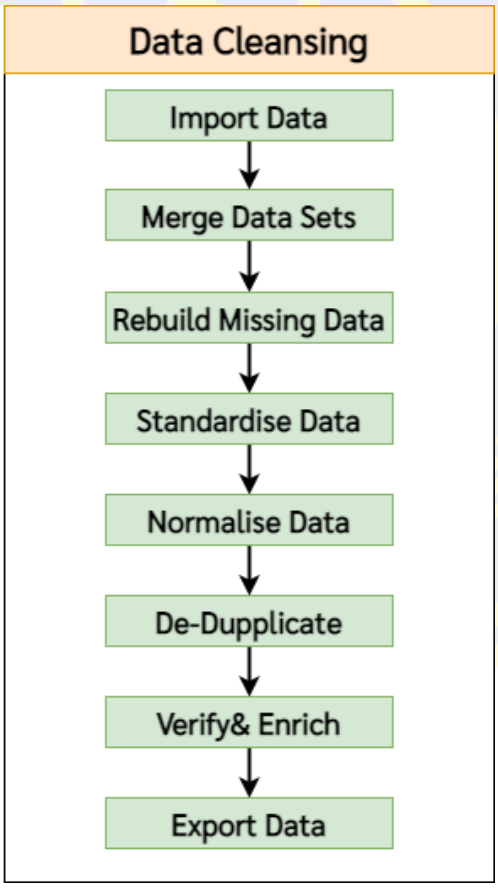


Figure 10 Data Cleansing

From phase data collection, it can be seen that the data received may come in a variety of formats, such as on paper, in PDF files, or via a website, the information may arrive in a variety of formats. Therefore, we must compile

these data into a single file. We've decided to keep it as an Excel file due of its usability.

A	B	C	D	E	F
time	tourists (person)	avg_income (bath)	population (person)	student_enrollments (person)	solid_waste (ton)
2011-01-01	892599	23007	43322	12922	2242
2011-02-01	892599	23007	43322	12922	2173
2011-03-01	892599	23007	43322	12922	2422
2011-04-01	892599	23007	43322	12922	2355
2011-05-01	892599	23007	43322	12922	2473
2011-06-01	892599	23007	43322	12922	2380
2011-07-01	892599	23007	43322	12922	2458
2011-08-01	892599	23007	43322	12922	2447

Figure 11 Cleaned data

Here is an example of the format of the data in the Excel file, which can be used for future reference. The data included all five factors and time.

2. Data transformation

   The data has been collected and is ready to be entered into an Excel spreadsheet. The data must be converted to CSV format before being imported into the model. This allows for easier usage and data manipulation in the next step.

Phase 3: Training and Testing

The dataset will be split into training and testing datasets. This is one of the most important aspects of the machine learning process. We utilized the sklearn library's function in this case. We split our dataset so that 80 percent of the data is in the training phase and 20 percent is in the testing phase. This ratio is selected due to our previous research, which suggested that the ratio of 80:20 delivers the acceptable results. As a result, feature scaling should always be done after the train-test split.

```
from sklearn.model_selection import train_test_split
train, test = train_test_split(df, train_size=0.8, random_state=7)
```

```
[ ]  1  train.head()
```

	tourists (person)	avg_income (bath)	population (person)	student_enrollments (person)	solid_waste (ton)
15	876303	25687	43304	14943	2122
11	892599	23007	43322	12922	2180
43	719644	27812	45142	12352	2669
50	779854	27257	46063	11397	3008
107	1232880	28706	46658	8416	3328

```
[ ]  1  test.head()
```

	tourists (person)	avg_income (bath)	population (person)	student_enrollments (person)	solid_waste (ton)
51	779854	27257	46063	11397	2828
84	1261842	28186	46352	7338	3103
65	1111247	27461	46302	11912	2859
93	1261842	28186	46352	7338	3737
46	719644	27812	45142	12352	2615

Figure 12 Example of training and testing data

Figure 12 illustrates a part of the data chosen at random from the proportionate separation of the data into test data and train data.

CHAPTER 4
RESULT

4.1 Overview of the Data

The table 5 displays the mean, median, and minimum to maximum values for each factor. There are 108 data entries, which correspond to 108 months. It shows overview of the collected data from 5 factors. The Tourists, Average Income, Population, and Student Enrollments are independent variables, whereas the Solid Waste is dependent variable. According to the algorithm requirements, these variables must have the same amount. Detailed figures of each factor can be obtained from table 5.

Table 5 Overview of the Data

Index	Tourists (person)	Average Income (baht)	Population (person)	Student Enrollments (person)	Solid Waste (ton)
count	108	108	108	108	108
mean	979903	27127	45225	11168	2843
std	213800	1677	1383	2743	425
min	719644	23007	43304	7087	2122
0.25	779854	27257	43457	8416	2514
0.5	892599	27665	46063	11912	2809
0.75	1217968	28186	46352	12922	3144
max	1261842	28706	46658	14943	3737

1. count: Count number of non-NA/null observations.
2. mean: Mean of the values.
3. std: Standard deviation of the observations.
4. min: Minimum of the values in the object.
5. 0.25: Lower percentile of the values in the object.
6. 0.5: 50 percentile is the same as the median.
7. 0.75: Upper percentile of the values in the object.
8. max: Maximum of the values in the object.

For numeric data, the result's index will include count, mean, std, min, max as well as lower, 50 and upper percentiles. By default, the lower percentile is 25 and the upper percentile is 75. The 50 percentile is the same as the median. Standard deviation (SD) is a common statistical distribution measure used to compare how widely the values in a data set are distributed. If most of the data is very close to the mean, then the standard deviation is small. On the other hand, if each data point is largely far from the mean, then the standard deviation is large. And, when all data has the same value The standard deviation or is equal to zero, i.e., there is no distribution. One useful feature is that the standard deviation uses the same units as the data.

4.2 Correlation

The Pearson Correlation (PC) technique is the most often used approach for numerical variables. It gives a number between -1 and 1, where 1 represents whole positive correlation, -1 represents total negative correlation, and 0 represents no correlation. It can be interpreted as follows: a correlation value of 0.70 between two variables indicates the existence of a substantial and positive association between the variables. A positive correlation indicates that if variable A rises, B will likewise increase, while a negative correlation indicates that if A increases, B will decrease.

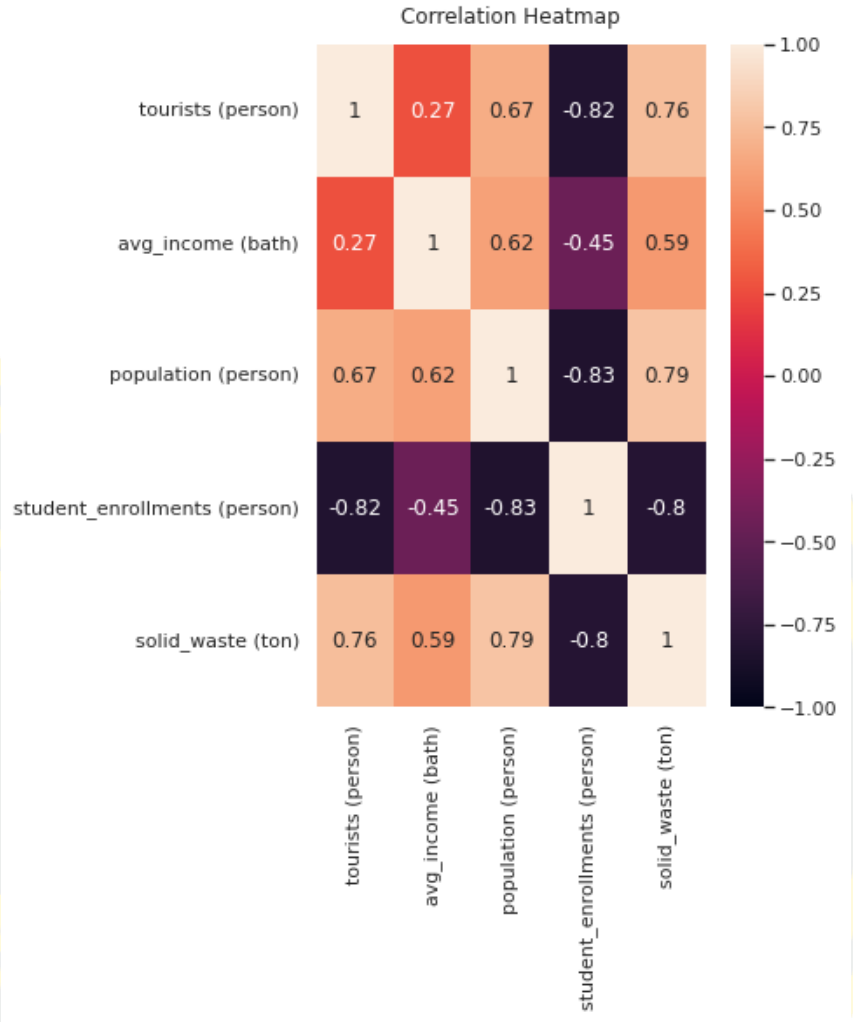


Figure 13 Correlation heatmap of each factors

The following set of interpretative criteria may be used to make sense of the data shown in the figure 13 of heatmap correlation up top.

Table 6 The correlation result of each studied factors

Correlation Coefficient	**Relationship Level**
0.71 - 1.00	highly relationship
0.51 - 0.70	moderate relationship
0.21 - 0.50	low relationship
0.01 - 0.20	very little relationship.
0.00	no relationship
-0.01 to -1	negative relationship

Table 6 represents the correlation result of each studied factors as a heatmap.

The highly correlated variables

- The relationship between population and amount of solid waste, and
- The relationship between tourists and the amount of solid waste.

The moderately correlated variables

- The relationship between population and tourists,
- The relationship between population and average monthly household expenses
- The relationship between amount of solid waste and average monthly household expenses.

The low correlated variable

- the relationship between population and average monthly household expenses.

The negatively correlated variable

- The relationship between the student enrollments with all 4 remaining variables.

4.3 Multiple Linear Regression (MLR)

In this process, the prepared data will be divided into 80% of training and 20% of testing data. Then, the dataset was staged seven times at random.

The Linear Model of a Linear Regression is imported from the sklearn library. It results with R-squared = 0.74623.

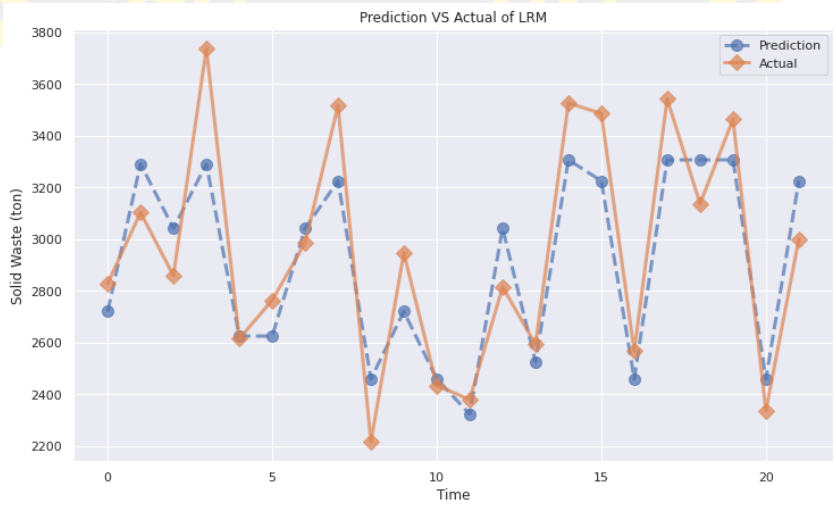


Figure 14 Result of Multiple Linear Regression

Figure 14 depicts the comparison of solid waste prediction and actual data.

- X is time
- Y is solid waste (Ton)
- The orange line (diamond marker) is actual data.
- The blue dashed line (circle marker) is prediction data.

According to Figure 14, the graphs are aligned properly and the R-squared results are satisfactory. This model can be utilized effectively.

4.4 Long Short-Term Memory (LSTM)

An LSTM is an algorithm in ANN category. The data was prepared and divided into 80% of training and 20% of testing data as before. In the modeling process, the neural network consists of an LSTM layer followed by a dense layer with a linear activation function. Then, compiling with a loss like Mean Square Error (MSE) and Adam optimizer. The model consists of 100 neurons and 400 calculation epochs. The evaluation metrics are final loss = 0.00292 and final validation loss = 0.00332.

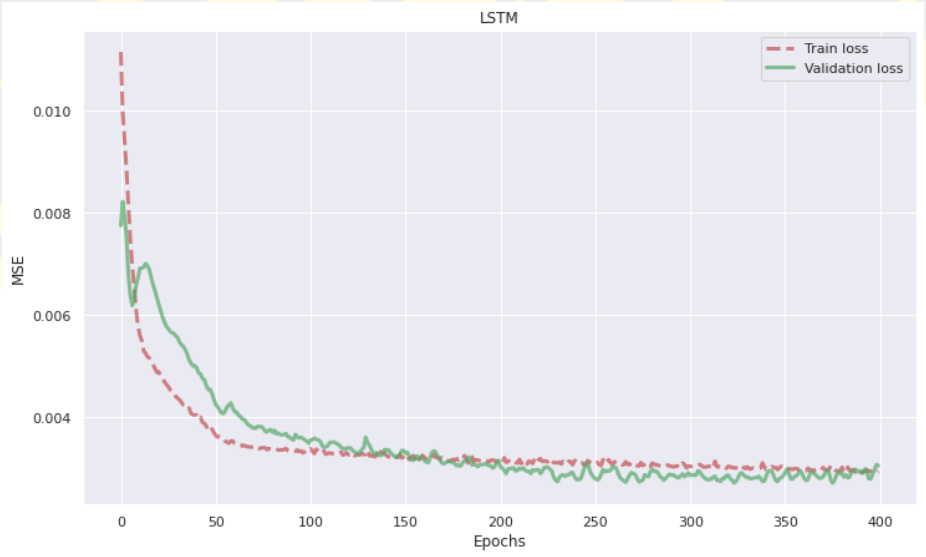


Figure 15 LSTM Training phase result

As depicted in the figure 15, the MSE values of the LSTM model were obtained after 400 training steps. The validation loss was continuously reduced and the results were satisfactory and can be continued. The results are MAE = 0.04154 and MSE = 0.00303. The interpretation is that the lower the value, the better the properties for use. Therefore,

we decided to use MSE to train the model and the R-squared = 0.57177, indicating that the values are satisfactory.

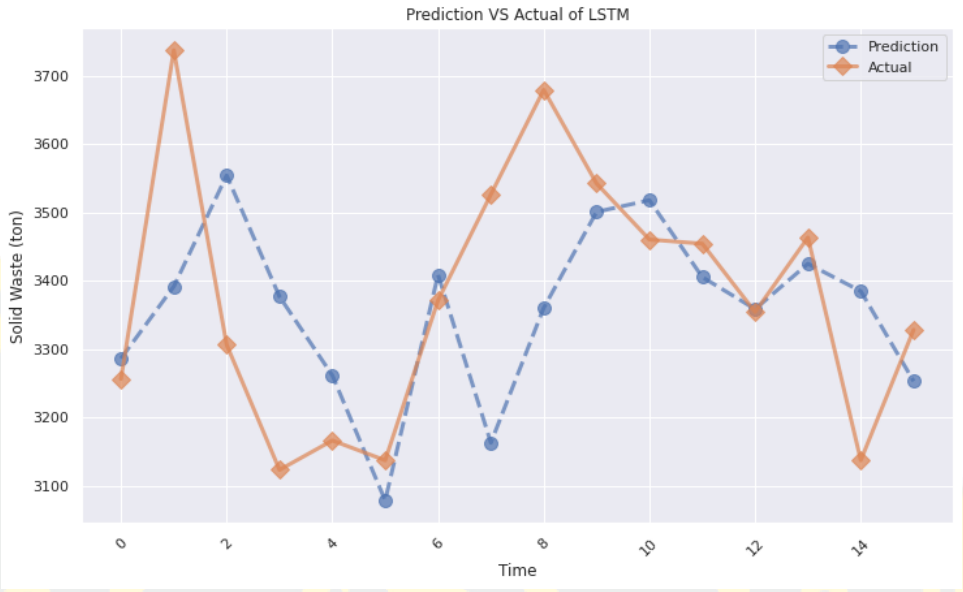


Figure 16 Result of Long Short-Term Memory

Figure 16 depicts the comparison of solid waste prediction and actual data.

- X is time
- Y is solid waste (Ton)
- The orange line (diamond marker) is actual data.
- The blue dashed line (circle marker) is prediction data.

According to figure 16, the graph's alignment may not be as satisfactory as it should be. This affects the interpretation of the R-squared value, which is not as satisfactory as it should be. This may indicate that this model is unsuitable for our dataset.

4.5 Support Vector Regression (SVR)

SVR is a regression function that is generalized by SVM. The non-linear data was prepared and divided into 80% of training and 20% of testing data as before and was staging seven times at random. Then, it is imported into a linear model of SVR function from sklearn library. We then passed it to a kernel called Radial Basis Function (RBF) kernel, which results in R-squared= 0.70785.

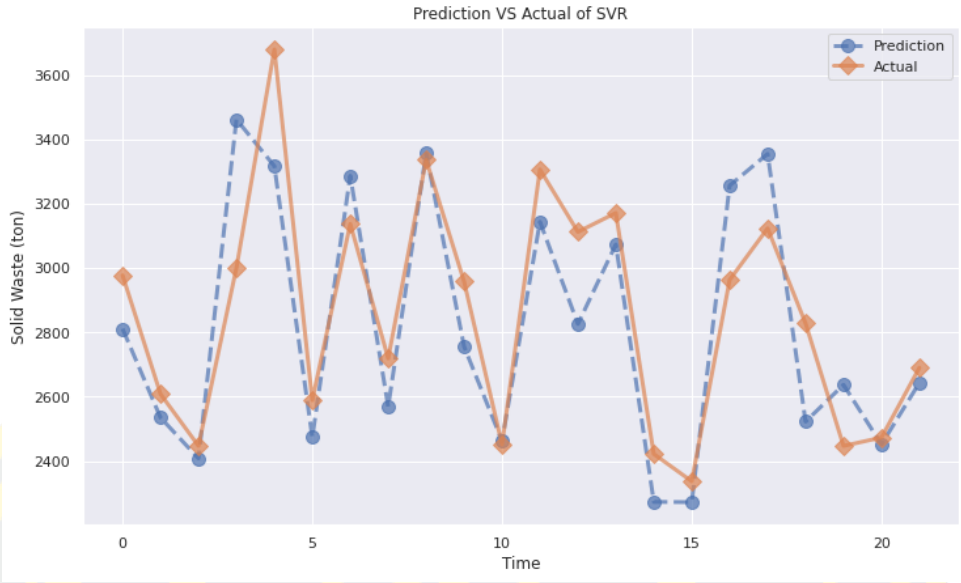


Figure 17 Result of Support Vector Regression

Figure 17 depicts the comparison of solid waste prediction and actual data.

- X is time
- Y is solid waste (Ton)
- The orange line (diamond marker) is actual data.
- The blue dashed line (circle marker) is prediction data.

According to figure 17, the graphs are aligned properly and the R-squared results are satisfactory. This model could be utilized effectively

CHAPTER 5
DISCUSSION AND CONCLUSIONS

Discussion

The collection of raw data from Seansuk Municipality is featured with various significant factors. These selected 5 factors are similar to the case studies, as well as the size of the city, the population, and the behavior and activities of the residents. The data was collected from sources existing in different departments of Saensuk Municipality. The data received may come in a variety of formats, such as on paper, in PDF files, or via a website, the information may arrive in a variety of formats. Therefore, we must compile these data into a single Excel file due to its usability.

This demonstrates the significance of data management and systematic dissemination. As a result of the necessity of storing and disseminating relevant knowledge so that it can be promptly and easily utilized for next purposes. This research is limited by the small amount of accessible data; however, if more data are collected, the model can be processed with more precision. As a result, we must pay closer attention to the storage, preservation, and diffusion of information, as it may be of considerable use in the future. The more frequently data is collected, the more effectively it may be processed to determine a result or correlation. From the data collection process, it was found that many factors were collected once a year. If there is monthly or weekly data, we may find important information that can be used to improve the municipality or organization.

This prediction model is based on existing data from the Saensuk Municipality. Unfortunately, this data is sufficient only for a rough prediction model. For more precise model development, we must have detailed data collecting from Saensuk Municipality. In this case, Saensuk Municipality has to collect data more accurately, e.g. the solid waste should be collected and reported daily, and the solid waste should be separated into recycle waste and non-recycle waste, etc.

Key Findings

We found that if more frequencies were collected and used as input to the model, the factors contributing to the increase in waste were better known. This model can be used

for solid waste generation planning during tourist high season or big events. The municipality can plan ahead in terms of waste collection trucks and bins. For example, if we know the number of tourists or other factors, for instance, and input those data into the model, the model can predict the amount of waste in that time. When municipal or local authorities are aware of the amount of waste that will be generated in the future, they can ensure that waste collection vehicles and bins are sufficient to accept waste from tourists. They can plan to raise or decrease the number of bins and rounds of waste collection to accommodate the approaching waste must not give either too many or too few. Ultimately, this will reduce excessive energy usage and associated expenditures.

Conclusion

In this research, we collected relevant existing data of solid waste management from the Saensuk Municipality to be used in the prediction model development. For finding state of the art, we reviewed the works between 2007 and 2021 related to the development of models for predicting the generation of solid waste. We analyzed and compared various models that had been applied and selected in that works. As a result, Linear Regression (LR), Artificial Neural Network (ANN), and Support Vector Machine (SVM) models are some of the most commonly used as predicting tools in solid waste management. In the data preparation process, the acquired data from Saensuk will be engineered and prepared. Then the prepared data was converted into a suitable format and loaded into the model for analysis. The data was fed into the models, which will build the prediction models and then return the model's results in training phase. Finally, the results of each model were compared to obtain the best model for predicting the generation of solid waste in Saensuk Municipality.

According to Pearson Correlation analysis, the population, the number of tourists, and the average monthly household expenses are influential variables for the Municipal Solid Waste (MSW) generation in Saensuk. We discovered that the Relatively Correlation coefficient of MLR, SVR, and LSTM are R-squared=0.74623, R-squared=0.70785, and R-squared=0.57177 respectively. The best R-squared result is 1.0. However, the Relatively Correlation coefficient alone may not enough to assure the correctness of the models. The analysis of graph comparison between prediction graph and actual data graph will give more information. The alignment of the blue and

orange line chart in figure 3, figure 5, and figure 6 should have similar pattern. The graph results suggest that the MLR model is superior to the SVR model, with LSTM model performs the worst. This allows us to determine that the MLR model is best suited to our dataset, which has constraints and factors locally. This model can be used to predict solid waste generation in Saensuk Municipality, Chon Buri. In this work, we investigated publications related to municipal solid waste prediction models. We then selected 3 most used algorithms in conjunction with the hyperlocal conditions and factors. We found that the MLR is the best model for our case study with the R-squared = 0.74623 and the graph analysis indicates that this model is appropriate for prediction. The next step is to implement the model at Saensuk Municipality, Chon Buri, to get the field study results. This solid waste generation prediction model for Saensuk Municipality is an emergence of future solid waste for this region. It will provide information to aid management decision-making. If the management is able to make the right decisions, it will help reduce the environmental impact in a sustainable

Future Work

Over the course of nine years of data collection, we have arranged it in an Excel-formatted Google Sheet that is accessible if made readily available online. And in the developed section of the model, these are the Google Sheet data to be handled online at the Google Collaboratory. If there is a model file that we have already developed, the user can run it and get the results of the analysis. Lastly, if the organization, local authority, or those interested in further development can upload the data to the Google Sheet file in accordance with the format we've designed, they can immediately receive the findings of the analysis. There are those who will continue to develop by adding other factors to test the model can be done as well. This model can be used for solid waste generation planning during tourist high season or big events. The municipality can plan ahead in terms of waste collection trucks and bins. In the future, for getting more accurate prediction model, we plan to examine with other factors and collect more data. In addition to solid waste, it can be applied to the type of waste that is recyclable.

มหาวิทยาลัยบูรพา
BURAPHA UNIVERSITY

REFERENCES

Colab, G. (2022). Welcome To Colaboratory. Retrieved from https://colab.research.google.com/#scrollTo=Nma_JWh-W-IF

Dissanayaka, D., & Vasanthapriyan, S. (2019). *Forecast municipal solid waste generation in Sri Lanka*. Paper presented at the 2019 international conference on advancements in computing (ICAC).

Elshaboury, N., Mohammed Abdelkader, E., Alfalah, G., & Al-Sakkaf, A. (2021). Predictive Analysis of Municipal Solid Waste Generation Using an Optimized Neural Network Model. *Processes*, *9*(11), 2045.

MathWorks. (2022). What Is a Linear Regression Model? Retrieved from https://www.mathworks.com/help/stats/what-is-linear-regression.html

Matplotlib. (2022). Matplotlib. Retrieved from https://matplotlib.org/

Meza, J. K. S., Yepes, D. O., Rodrigo-Ilarri, J., & Cassiraga, E. (2019). Predictive analysis of urban waste generation for the city of Bogotá, Colombia, through the implementation of decision trees-based machine learning, support vector machines and artificial neural networks. *Heliyon*, *5*(11), e02810.

Nguyen, X. C., Nguyen, T. T. H., La, D. D., Kumar, G., Rene, E. R., Nguyen, D. D., . . . Nguyen, V. K. (2021). Development of machine learning-based models to forecast solid waste generation in residential areas: A case study from Vietnam. *Resources, Conservation and Recycling*, *167*, 105381.

NumFOCUS, I. (2022). Pandas. Retrieved from https://pandas.pydata.org/

NumPy. (2022). NumPy. Retrieved from https://numpy.org/

Office, N. S. (2022). 8 Revenue and Household expenditure Branch. Retrieved from http://statbbi.nso.go.th/staticreport/page/sector/th/08.aspx

Office, S. M. (2022a). History Saensuk Municipal. Retrieved from https://www.saensukcity.go.th/en/component/content/category/8-about.html

Office, S. M. (2022b). Population Statistics. Retrieved from https://saensukcity.go.th/population-statistics.html

Registrar, O. o. t. (2022). Number of new students. Retrieved from https://reg.buu.ac.th/registrar/stat.asp?avs165703731=2

Scikit-learn. (2022a). Scikit-learn. Retrieved from https://scikit-learn.org/stable/

Scikit-learn. (2022b). Support vector machines. Retrieved from https://scikit-learn.org/stable/modules/svm.html

Soni, U., Roy, A., Verma, A., & Jain, V. (2019). Forecasting municipal solid waste generation using artificial intelligence models—a case study in India. *SN Applied Sciences*, *1*(2), 1-10.

Sports, M. o. T. (2022). Domestic Tourism Statistics Q1-Q4 (Classify by region and province). Retrieved from https://www.mots.go.th/more_news_new.php?cid=618

Sun, N., & Chungpaibulpatana, S. (2017). Development of an appropriate model for forecasting municipal solid waste generation in Bangkok. *Energy Procedia*, *138*, 907-912.

Waskom, M. (2022). An introduction to seaborn. Retrieved from https://seaborn.pydata.org/tutorial.html

Wikipedia. (2022a). Artificial neural network. Retrieved from https://en.wikipedia.org/wiki/Artificial_neural_network

Wikipedia. (2022b). Pearson correlation coefficient. Retrieved from https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
Wikipedia. (2022c). Principal component analysis. Retrieved from https://en.wikipedia.org/wiki/Principal_component_analysis
Wikipedia. (2022d). Random forest. Retrieved from https://en.wikipedia.org/wiki/Random_forest

APPENDICES

Implementation of a model using Google Collaboratory

Google Colaboratory

Google Colaboratory is a Google Research product. Colab enables anyone to create and execute arbitrary Python code via a web browser and is particularly suitable for machine learning, data analysis, and teaching. Colab is technically a hosted Jupyter notebook service that requires no configuration and provides free access to computational resources, including GPUs (Colab, 2022) .

Connect Google Colab with Google Drive

1. Open the Google Drive
2. Search for Colab in the Google Workspace Marketplace
3. Install the Colaboratory.

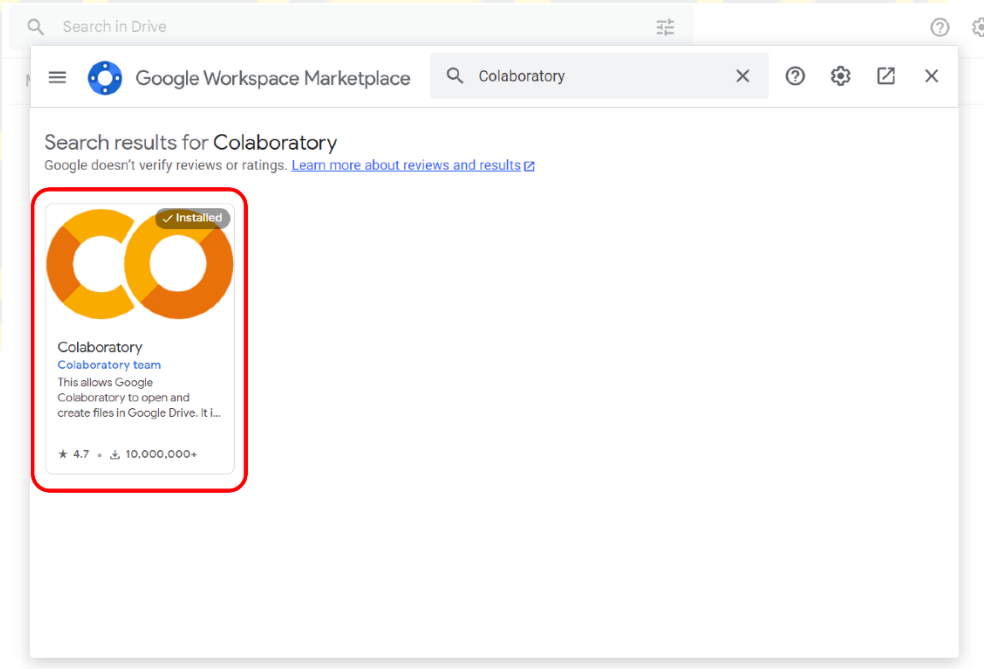


Figure 18 Install the Colaboratory

4. Create a new file by right-clicking Google Drive
5. Select menu “More”
6. Select menu “Google Colaboratory”.

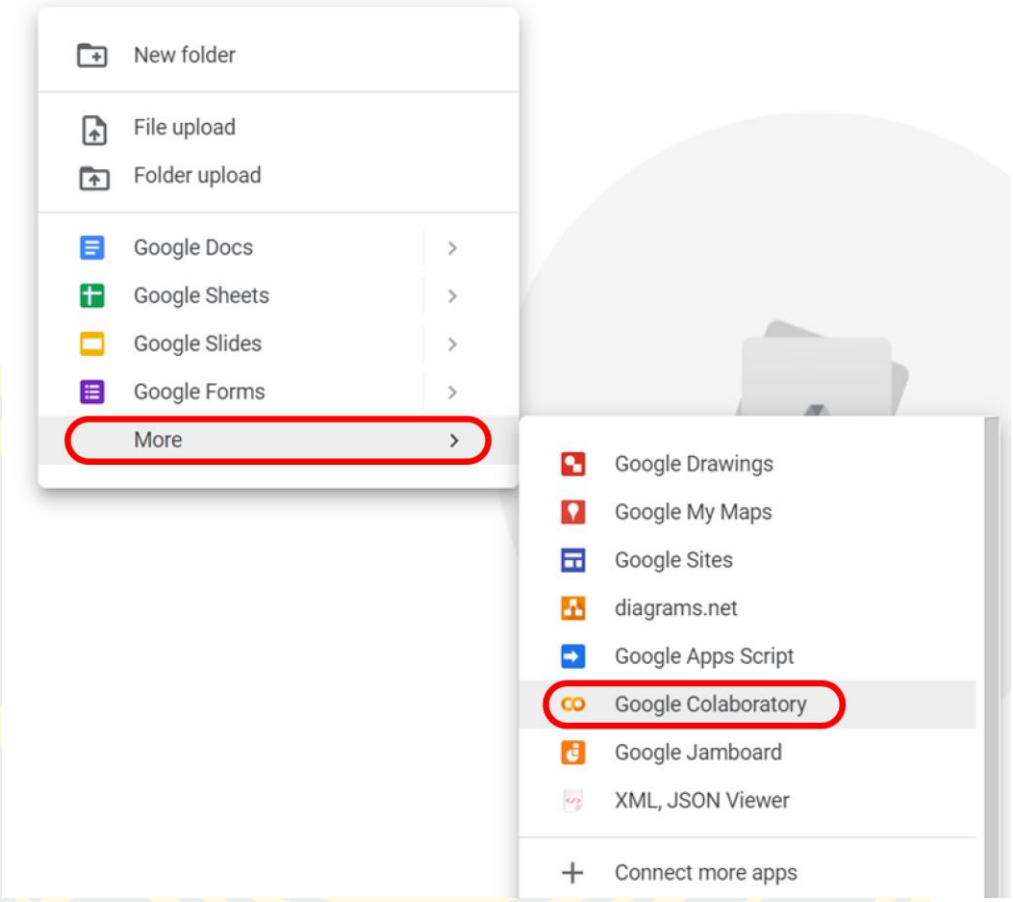


Figure 19 New colaboratory file

7. Enter this code to import drive here and mount them together.

```
from google.colab import drive
drive.mount('/content/drive')
```

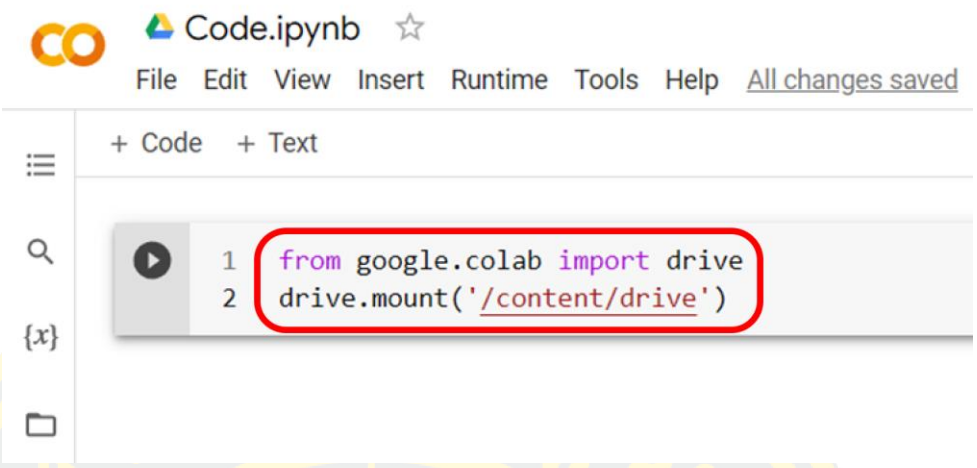


Figure 20 Mounting file

8. Run the following script in colab shell and click menu “Runtime” and “Run all” or use the keyboard shortcut, press and hold either Ctrl key, and while continuing to hold, press F9

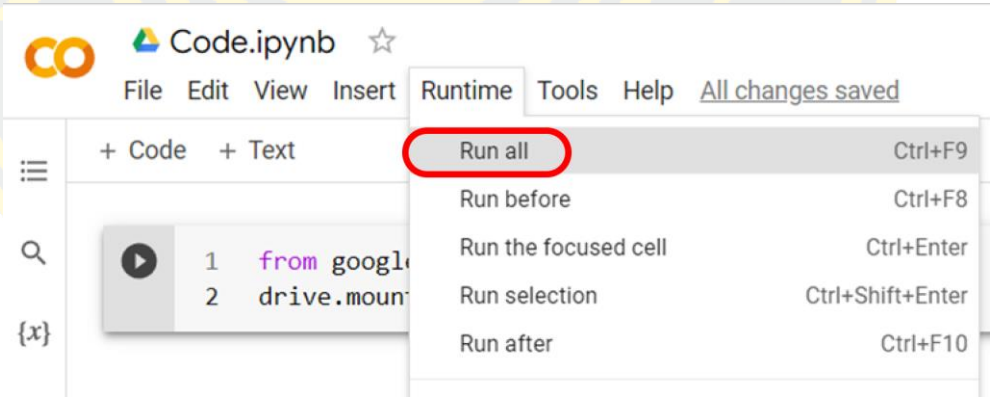


Figure 21 Run file colab

9. To allow the notebook access to our Google Drive files, click the button labeled “Connect to Google Drive”.

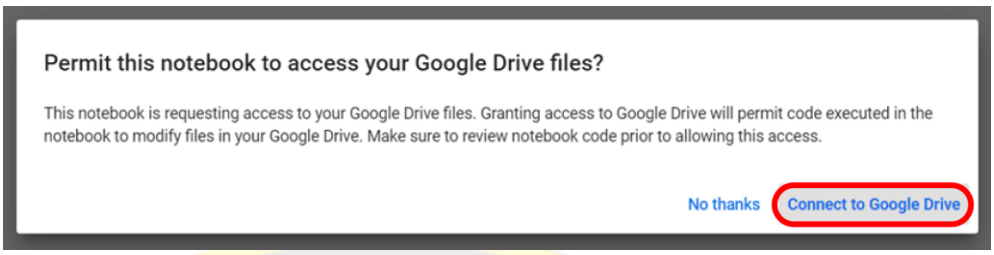


Figure 22 Connect to Google Drive

10. Peruse the list of requests for access permissions. Then, click the "Allow" button.

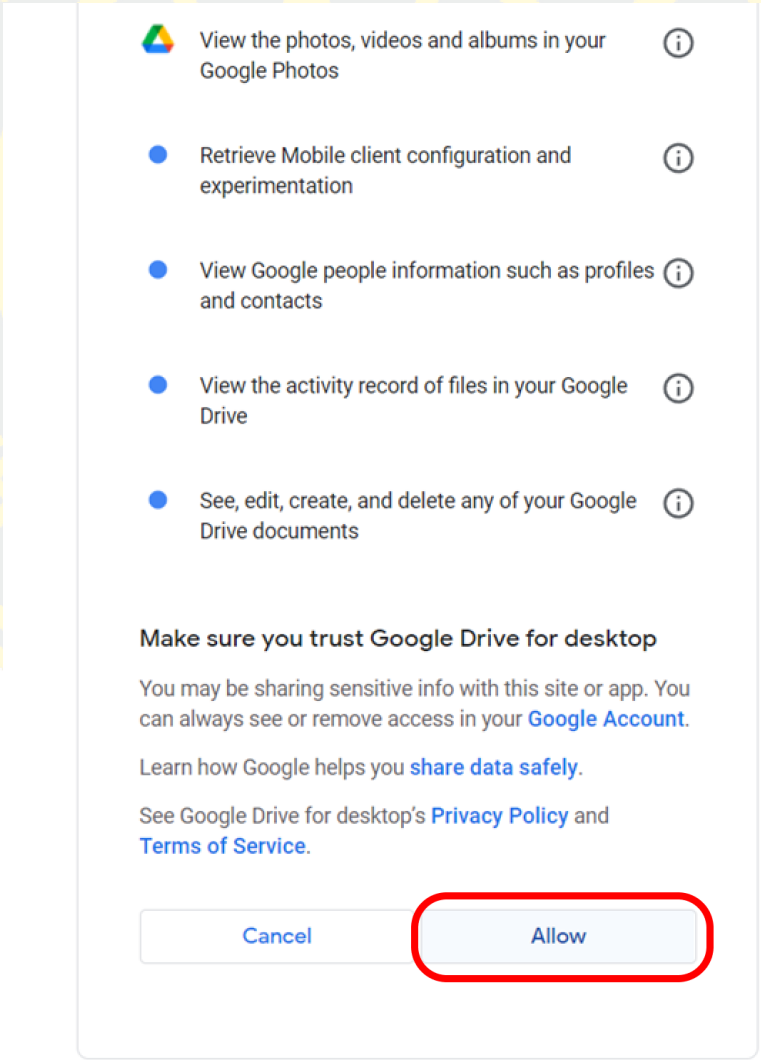
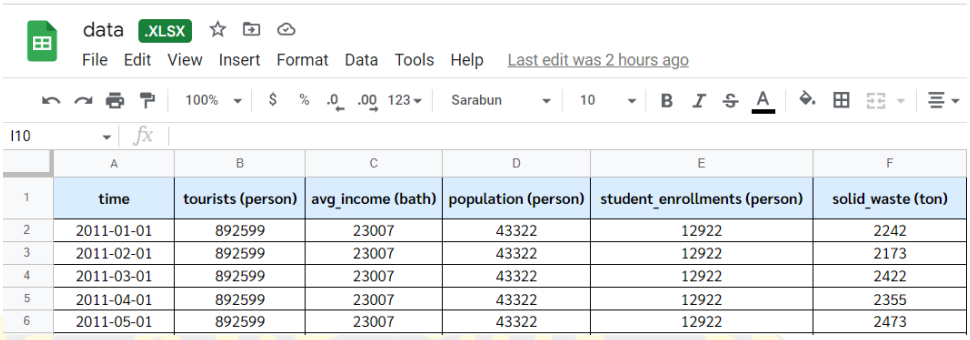


Figure 23 Requests for access permissions

11. Create or import an Excel spreadsheet and save it to Google Drive.



	A	B	C	D	E	F
1	time	tourists (person)	avg_income (bath)	population (person)	student_enrollments (person)	solid_waste (ton)
2	2011-01-01	892599	23007	43322	12922	2242
3	2011-02-01	892599	23007	43322	12922	2173
4	2011-03-01	892599	23007	43322	12922	2422
5	2011-04-01	892599	23007	43322	12922	2355
6	2011-05-01	892599	23007	43322	12922	2473

Figure 24 Example data in google sheet

12. Enter this code to read a Excel file form Google Drive.

```
df = pd.read_excel('/content/drive/MyDrive/data.xlsx')
df
```

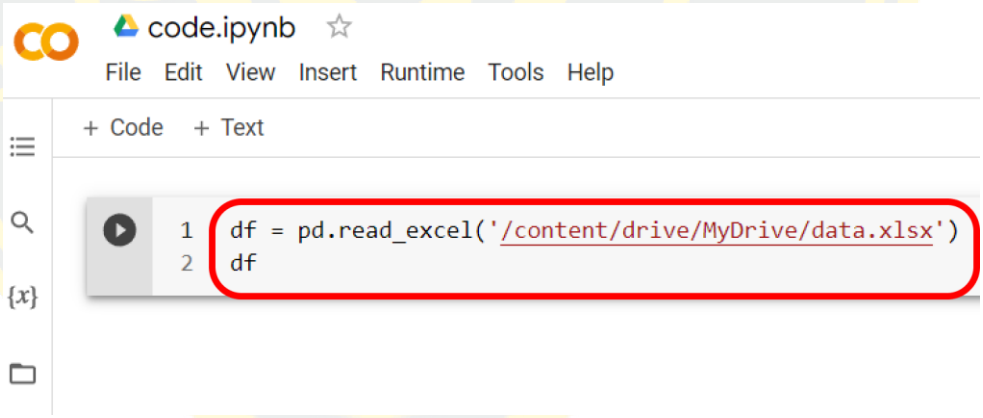


Figure 25 Reading data from google sheet file

13. Run the following script in colab shell and click menu “Runtime” and “Run all” or use the keyboard shortcut, press and hold either Ctrl key, and while continuing to hold, press F9. The results will be displayed below.

	time	tourists (person)	avg_income (bath)	population (person)	student_enrollments (person)	solid_waste (ton)
0	2011-01-01	892599	23007	43322	12922	2242
1	2011-02-01	892599	23007	43322	12922	2173
2	2011-03-01	892599	23007	43322	12922	2422
3	2011-04-01	892599	23007	43322	12922	2355
4	2011-05-01	892599	23007	43322	12922	2473
...	...	...	...	...	...	...
103	2019-08-01	1232880	28706	46658	8416	3454
104	2019-09-01	1232880	28706	46658	8416	3354
105	2019-10-01	1232880	28706	46658	8416	3463
106	2019-11-01	1232880	28706	46658	8416	3137
107	2019-12-01	1232880	28706	46658	8416	3328

108 rows × 6 columns

Figure 26 Result data

Link of model and dataset

1. Model Multiple Linear Regression (MLR) https://drive.google.com/file/d/1mQ5GJttV94JGFFEPKE7VsAp1UCVj2QEb/view?usp=share_link
2. Model Long Short-Term Memory (LSTM) https://drive.google.com/file/d/16CkixDkXibclkSBPH9howEDLZc6NGAjG/view?usp=share_link
3. Model Support Vector Regression (SVR) https://drive.google.com/file/d/1XNDGEVUY6zI_Nrogjl11LOJwoXBttadw/view?usp=share_link
4. Dataset 1 https://docs.google.com/spreadsheets/d/1hSs2qICKoPh0EqWMzduXY8bXu-0nBvDc/edit?usp=share_link&ouid=101471441107704523469&rtpof=true&sd=true
5. Dataset 2 https://docs.google.com/spreadsheets/d/1gDYfhpF9YU7vB_emGdNy4aMc894dKCSh/edit?usp=share_link&ouid=101471441107704523469&rtpof=true&sd=true

BIOGRAPHY

NAME	Kittiya Thibuy
DATE OF BIRTH	07 May 1998
PLACE OF BIRTH	Chanthaburi
PRESENT ADDRESS	5/12 Soi2 , Bangsaen Sai 4 South Road, Saensuk, Muang Chonburi, Chonburi 20130
EDUCATION	Bachelor's degree: Informatics, Information Technology, Burapha University Master's degree: Data Science, Information Technology, Burapha University
AWARDS OR GRANTS	The Twenty-Second National Software Contest (NSC 2020) October-2019 - February-2020 Competition on special topics in the type of IoT in the project “Trash2Cash”
	The Computing Technology Industry Association (CompTIA) May-2019 - May-2019 Successfully completed the requirements to be recognized as CompTIA Cloud Essentials
	The 7th ASEAN Undergraduate Conference in Computing March-2019 - March-2019 Presented a research paper "Rewarding Smart Recycle Bin Prototype Designed for Saensuk Sub-District" with the award of Excellent Paper
	The Twenty-First National Software Contest (NSC 2019) March-2019 - March-2019 Honorable mention on special topics in the type of IoT in the project "Rewarding Smart Recycle Bin"
	Eastern Code Festival July-2018 - July-2018 The third prize of the Code Marathon competition in the IOS group