DS702: Programming for Data Science

Project Name: Thailand monthly Dengue patient estimation and visualization from weather data using simple linear regression

Team Name: What to Do If You Are Sick

Team Members:

1. Nuttawut Thuayhanruksa	Student ID: 640631030	Team Role: Visualization
2. Nattawat Wattanawiput	Student ID: 640631031	Team Role: Regression Model
3. Sonram Sirirat	Student ID: 640631037	Team Role: Data collection
4. Akkapop Prasompon	Student ID: 640631128	Team Role: Data preparation

Project Abstract:

Thailand is a tropical country that is suitable for the growth and reproduction of Aedes which is a carrier of dengue disease. An appropriate climate such as high rainfall rate and high temperature in Thailand might indicate the increasing of an Aedes reproduction rate that raises a chance of dengue fever spreading. Therefore, the objective of this project is to estimate the number of dengue patients that are related to the weather in Thailand such as temperature, rainfall, humidity value in each month by using temperature, rainfall, humidity data from the National Statistical Office of Thailand by linear regression model. To separate the patient group into the region by using k-means clustering, those regions are determined by climate. The risk groups are clustered by using dengue patients' data from the department of disease and visualizing the information into the chart using a python library such as Plotly. We use Polynomial regression with degree is equal to 2 to predict the number of patient's R-square score for the training set is 83% and R-square for the testing set is 78% with RMSE = 753.68, MAE = 544.95. For clustering, we cluster 3 groups by amount of rainfall (Low, Medium, High). Circle radius will show the patient count from each group.

1) Project Introduction:

Thailand is located in Southeast Asia which is a tropical climate area. Since the environment in Thailand is appropriate for mosquito reproduction. There is a famous breed of mosquito that can be found in Thailand called Aedes, Aedes is a carrier of a lethal disease named “dengue fever” [1] in humans. Some research found that the amount of rainfall measured is significantly associated with the number of dengue patients [2].

In Thailand, there are many organizations that work with data analysis tasks to predict and prepare for the incoming disaster such as the Meteorological Department of Thailand that predicts a chance of a thunderstorm that might occur in a country area and announce it to the people in the form of data visualization. Similar to our work, the objective of this project is to predict the number of dengue fever patients in Thailand from yearly data since 2017 using Thailand weather data such as average/minimum/maximum temperature, humidity, and measured rainfall value. As well, we designed to visualize the groups of patients into three specific groups including a high rain amount, medium rain amount and low rain amount.

Python is the main language for developing the prediction model and clustering model in this project. We designed to use standard packages of python such as pandas and NumPy for the data preparation process, Scikit-learn for fitting and evaluating the model, and Matplotlib for visualizing the result of predicted data. The weather data that we’re going to use is retrieved from the National Statistical Office of Thailand [3], and the patient data from the Department of disease control [4], also the Open Government Data of Thailand [5]. Selecting the best function from Scikit-learn library is the tool to select the features which should be a predictor.

The process to determine the prognosis of dengue fever is using the data from the number of dengue patients, humidity, rainfall, and temperature data via the regression model. We expect the accuracy rate of the model should be greater than 60%. And the k-mean clustering model should be applied. We also use a data visualization technique to specify which areas are likely to be at risk of dengue fever by using matplotlib. As well, the value of R2, MAE, MSE, and RMSE will be reported in the document. We believe that the result of our project will make the associated organization concerned about incoming problems that will lead to the decrease of dengue patients in Thailand.

2) Project Objectives:

- To predict the number of dengue fever patients by average rainfall, temperature, humidity.
- To cluster the dengue patient into multiple groups of regions.
- To summarize the result of prediction into an understandable form such as a report and data visualization.

3) Project Scopes:

1. The data that we will use to predict the number of dengue patients includes rainfall, temperature, and humidity data in Thailand between 2016 and 2020.
2. The main language that is used to develop the model is python 3.
3. A library such as scikit-learn, pandas, and NumPy will be implemented for development purposes.
4. A library such as matplotlib, and plotly will be implemented for data report and visualization purposes.

4) Project Details:

4.1 Project Overview Diagram

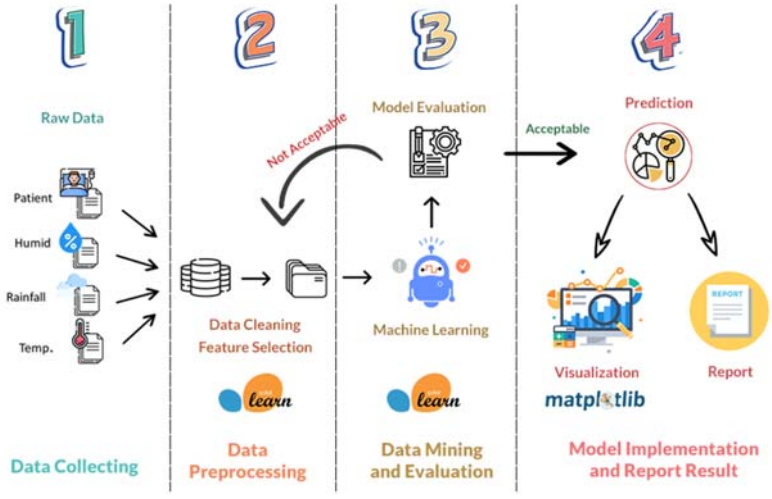


Figure 1: A diagram showing the overview of project development process

Above figure showing the whole process of project development (Figure 1). Procedure of project development is beginning from data preparation process that involve with data collecting from several sources, and data preprocessing which includes data cleaning methods such as remove missing value, imputation, and categorical feature encoding, also select the variable that is associated with the number of infected patients then split the data into training set and testing set. After all data has been prepared, we will move to the data mining process. In this process, we will keep fitting and evaluating the model using regression teach and cross-val-score function from scikit-learn library until the accuracy meets our condition. The testing data set will be implemented in the most optimal model and the result of the prediction such as R2, MAE, MSE, and RMSE will be reported, and the predicted result will be visualized using matplotlib library.

4.2 Project details

4.2.1 Data collecting

The government provides many datasets as open data. Not all of the data is available on the front page, such as patient data. Fortunately, the patient data they provide with no condition or any concerned policy to use also can't identify a person or sensitive information to track back who it is.

Patient dataset

Provider:	Department of Disease Control, Ministry of Public Health.
Source:	https://dvis2.ddc.moph.go.th/vizql/t/production/vudcsv/sessions/9BF277F0E90143D98BB4E0A1409ADB90-1:3/views/17214822742938834423_14405402496463851563?summary=true

Humidity, Rainfall, Temperature dataset

Provider:	National Statistical Office, Ministry of Digital Economy and Society.
Source:	https://osstat.nso.go.th/statv5/list.php?id_branch=21&Page=13

4.2.2 Data Preprocessing

Features selection

Feature name	Describe
year_num	code of year 2016 – 2020 (1-5)
province_num	code of province 1- 77
day_raindrop	how many days of rain drop in 365 days
quant_rain	total quantity rainwater in unit of millimeters
humbidity_perc	average % humidity in each province, each year
temp_max	maximum temp in each province, each year
temp_min	minimum temp in each province, each year
temp_avg	average temp in each province, each year
dead	amount of dead person from dengue
dead_perc	percent of dead person from dengue
patient	amount of patient from dengue (target variable)

Data cleaning

As the Thai dataset in some column(feature) names are in Thai language, we renamed them to the English language to avoid unexpected errors. The value in each column will be checked for null value before implemented to the model.

Data preparation

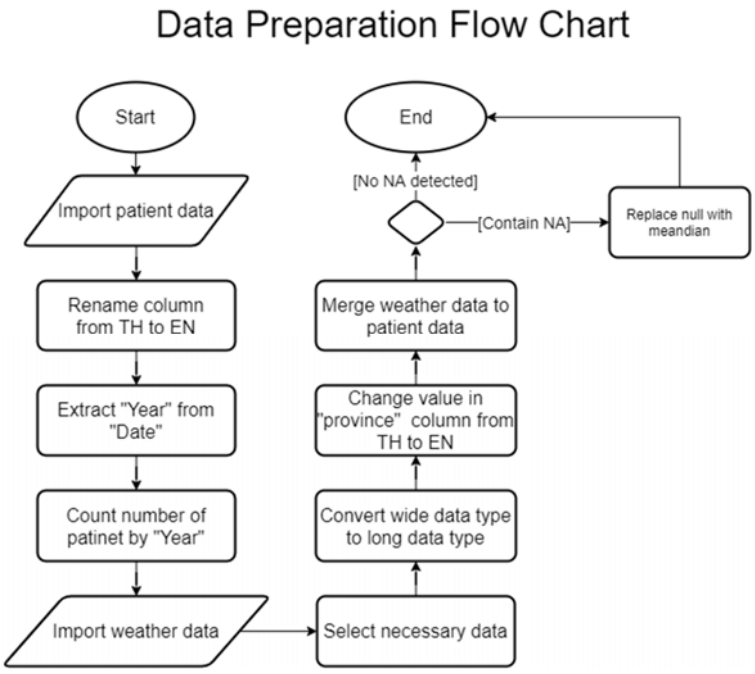


Figure 2: Flow chart diagram of the dengue data preparation process

Before training the model, the input data needs to be ensured that it won't cause any error to the algorithm. To accomplish that objective, the data has to be cleaned and preprocessed. The figure above (figure 2) shows the process of data preparation in this project. Each record in the patient data represents the data of a single patient who was infected from 2016 to 2021. Since we want to use only year data, we extracted the year from the date feature. The patient data from the source is represented case by case, to acquire the number of patients by year, we designed to use group_by method in order to summation the number of patients classified by "year", "province", and "age group".

After finishing the data preparation in patient data, we cleaned the weather data with the same steps as the patient data. The weather data contains humidity, temperature, and rain data. The column of this data set represents the year that they collected the data, the row represents the provinces, and data in the cell represents the value of each weather condition. We convert the wide data type to the long data type in order to make each row contain the data of each year. We also convert the province name from Thai to English by merging our dataframe with the dictionary data. The year has been converted from B.E. to A.D. by subtracting the year with 543. Finally, the patient data and weather data have been merged together by province and year and ready to be implemented in the next process.

4.2.3 Data Mining and Evaluation

We used the Linear Regression approach to create a model prediction for a dengue patient. The main libraries that we have used to develop a model are numpy, pandas, and scikit-learn. The seaborn and matplotlib library are also used for visualizing the data set and the result of prediction.

Model Testing: Linear Regression Baseline

1. Multiple Linear Regression
2. Repeats K-folds Cross Validation
3. Features selection: Recursive Feature Elimination (RFE)
4. Features selection: Recursive Feature Elimination with Grid Search CV
5. Polynomial regression

1. Multiple Linear Regression

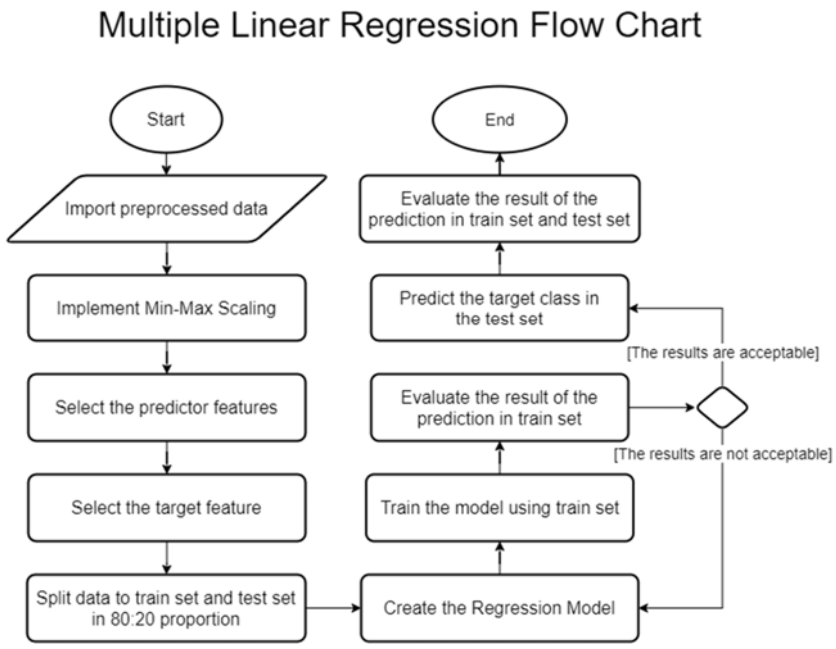


Figure 3: Flow chart diagram of building multiple linear regression model

Result of testing

After we applied simple model linear regression to the dataset, we have an R-squared score of 0.57 in the train set and 0.49 in the test set, Root Mean Squared Error (RMSE) of 812.8884 in our prediction result on this dataset. The overall Regression Evaluation Metrics is still less than expected then we will find a new method to make it more efficient.

2. Repeats k-folds Cross-Validation

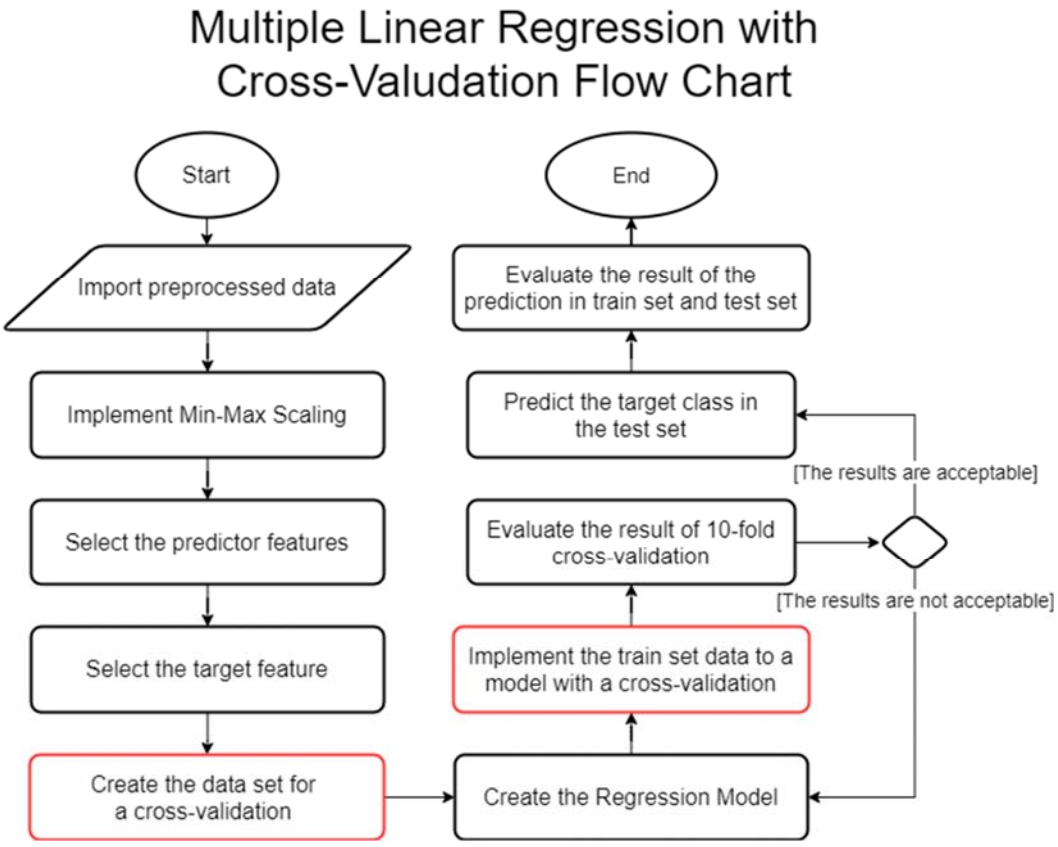


Figure 4: Flow chart diagram of building multiple linear regression model with a cross-validation

Result of testing

After we applied linear regression with Repeats k-folds cross-validation by a number of folders are 5 folds and a number of repeats are 3. We have an R-squared score of 0.57 in the train set and 0.52 in the test set, Root Mean Squared Error (RMSE) of 975.68 in our prediction result on this dataset. The overall Regression Evaluation Metrics is still less than expected and RMSE also increased, so we will find a new method to make it more efficient.

3. Features selection: Recursive Feature Elimination (RFE)

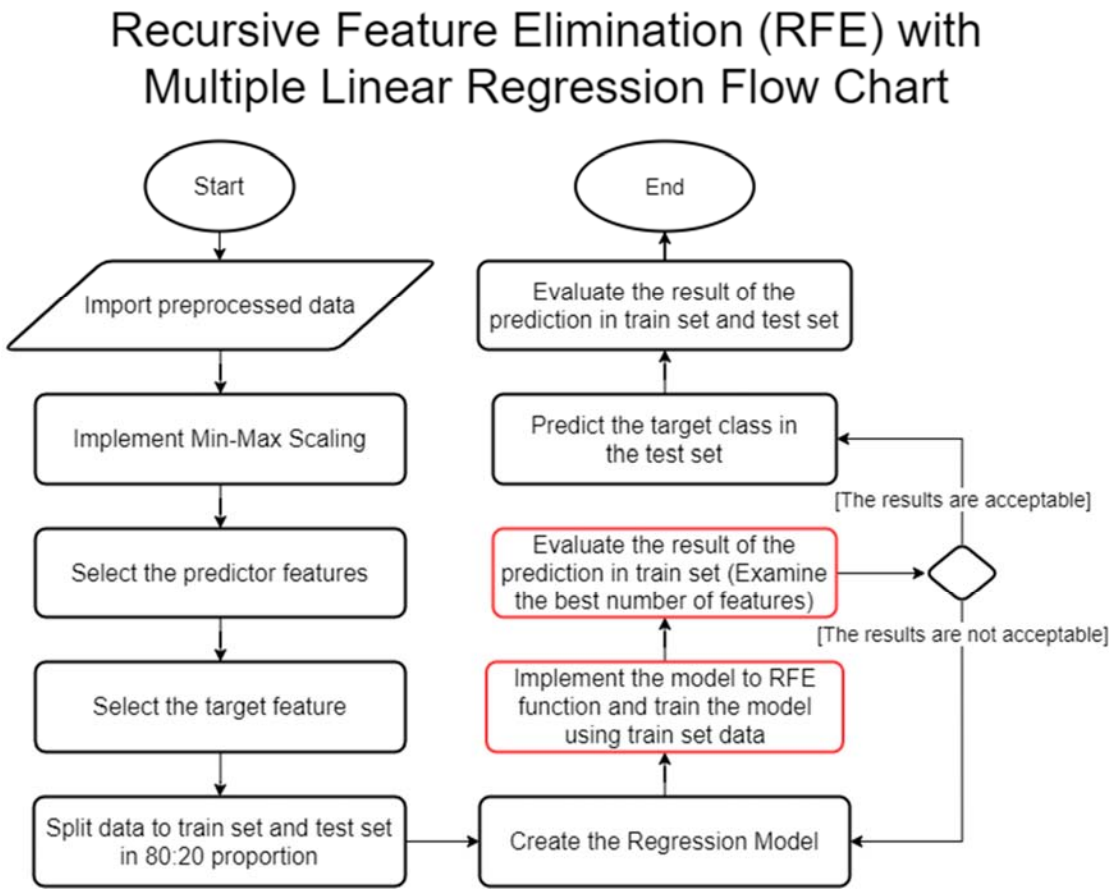


Figure 5: Flow chart diagram of building linear regression model and implement RFE function

Result of testing

After we applied linear regression that uses the Recursive Feature Elimination method (RFE). We have an R-squared score of 0.5634 in the train set and 0.5163 in the test set, Root Mean Squared Error (RMSE) of 790.60 in our prediction result on this dataset. The RFE got an optimal solution, the model dropped 5 columns that 1. Percent of humidity, 2. Rain quality, 3. Raindrop Day/year, 4. Year number and 5. Province number. If we select this model to deploy it will oppose our objective of the project, so we will find a new method.

4. Grid Search CV

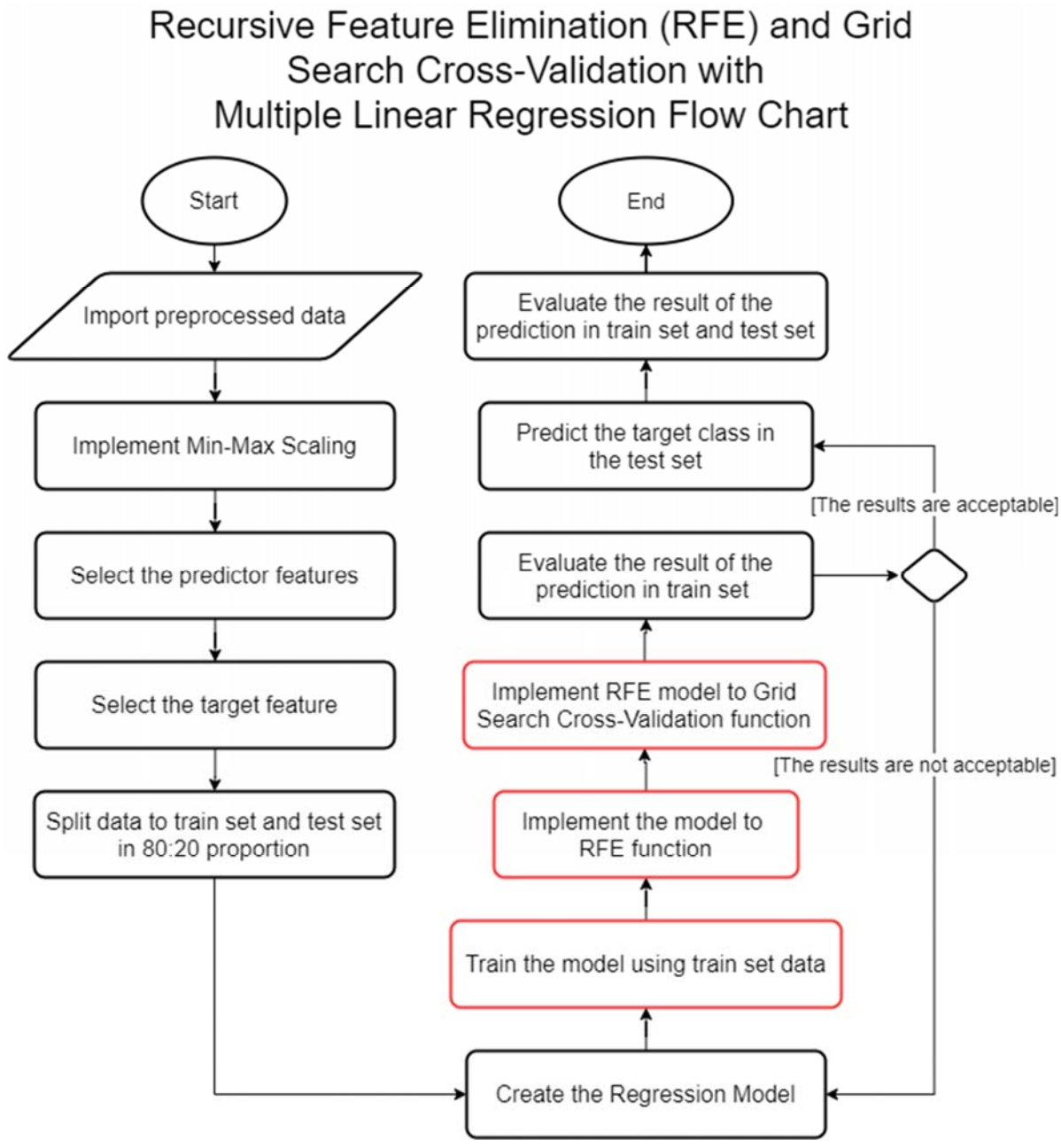


Figure 6: Flow chart diagram of building linear regression model and implement RFE model then applying a cross validation

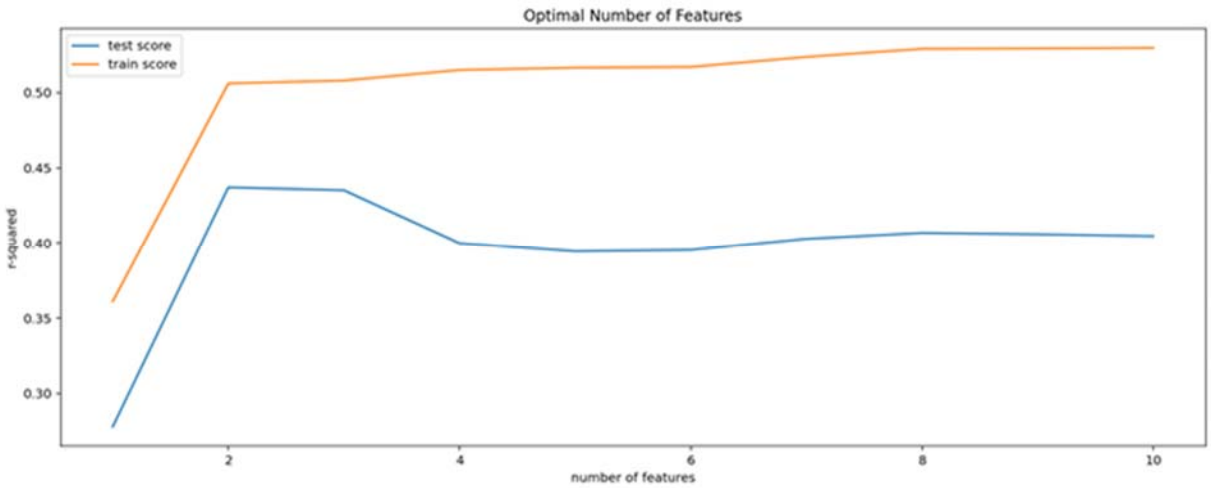


Figure 7: Line chart of the optimal number of feature after implemented RFE

Result of testing

After we applied linear regression that use the Grid Search CV and set up hyperparameters to tune

- Estimator = rfe (linear regression with backward method)
- Param_grid = hyper_params (features selection 1-10)
- Scoring= r2 (Coefficient of determination (R2 score))
- Cv = folds (use 5 folds)
- Verbose = 1 (use default value)
- Return_train_score=True

We have an R-squared score of 0.55 in the train set and 0.55 in the test set, Root Mean Squared Error (RMSE) of 864.66 in our prediction result on this dataset. The Grid Search CV got an optimal solution, they dropped 8 columns which make our hypothesis of our project fail. If we select this model to deploy, it will oppose our project's objective, so we will find a new method.

5. Polynomial regression (use degree = 2)

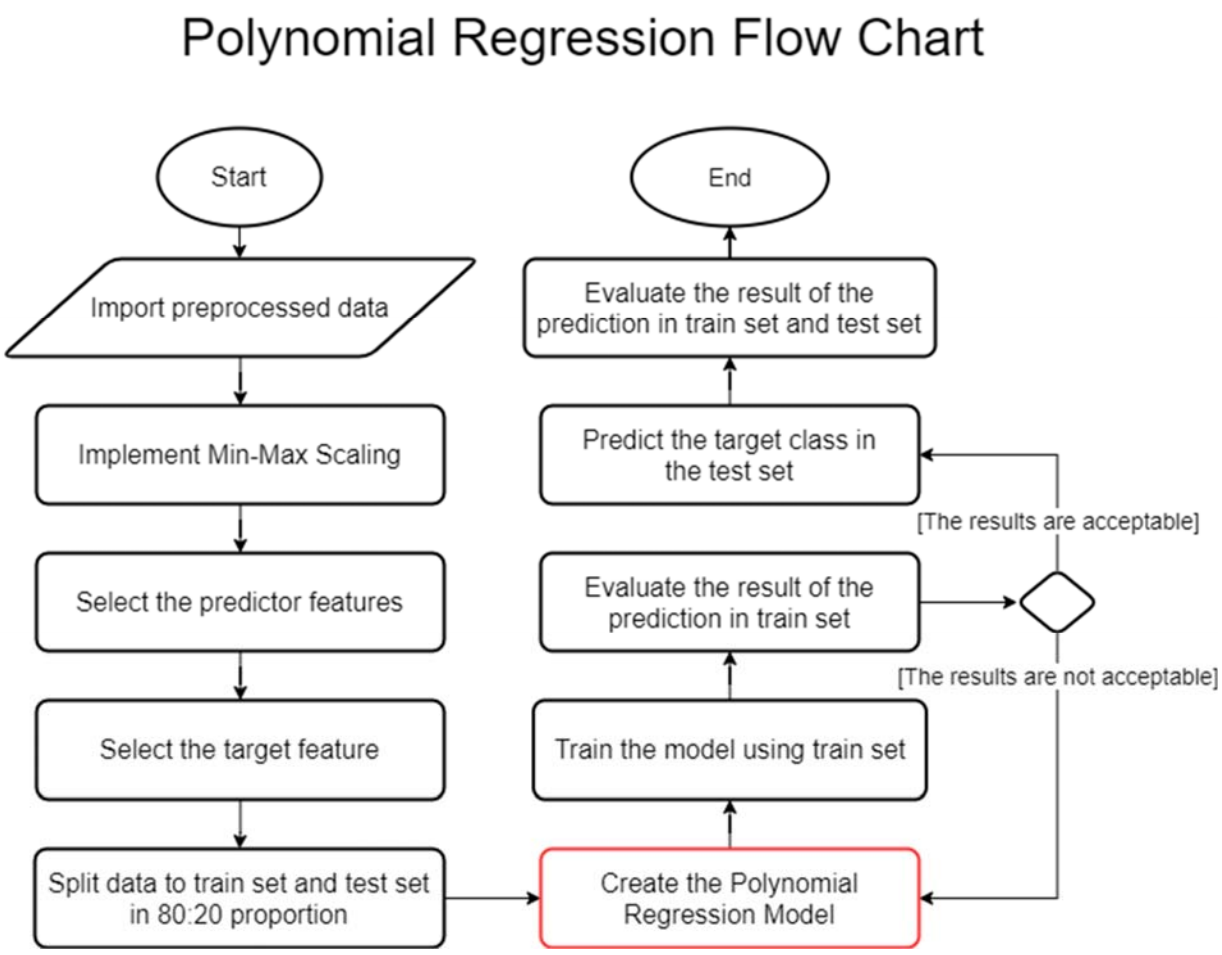


Figure 8: Flow chart diagram of building polynomial regression model

Result of testing

After we applied simple model polynomial regression to the dataset, we have an R-squared score of 0.83 in the train set and 0.78 in the test set, Root Mean Squared Error (RMSE) of 753.65 in our prediction result on this dataset. The overall regression evaluation metrics of Polynomial got the best result compared with 4 methods before.

Evaluation metrics

Coefficient of determination (R2 score) is used to evaluate the performance of a linear regression model. It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable(s). It is used to check how well-observed results are reproduced by the model, depending on the ratio of total deviation of results described by the model.

$$R^2 = 1 - \frac{\Sigma_i (y_i - f_i)^2}{\Sigma_i (y_i - \bar{y})^2}$$

Mean Absolute Error (MAE): the mean of the absolute value of the errors.

$$\frac{1}{n} \sum_{i=1} |y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE): the square root of the mean of the squared errors.

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

Comparison:

- R2 Score of Training set
- R2 Score of Testing set
- MAE is the easiest to understand because it's the average error.
- RMSE is even more popular than MSE, because RMSE is interpretable in the “y” units.

K-mean clustering

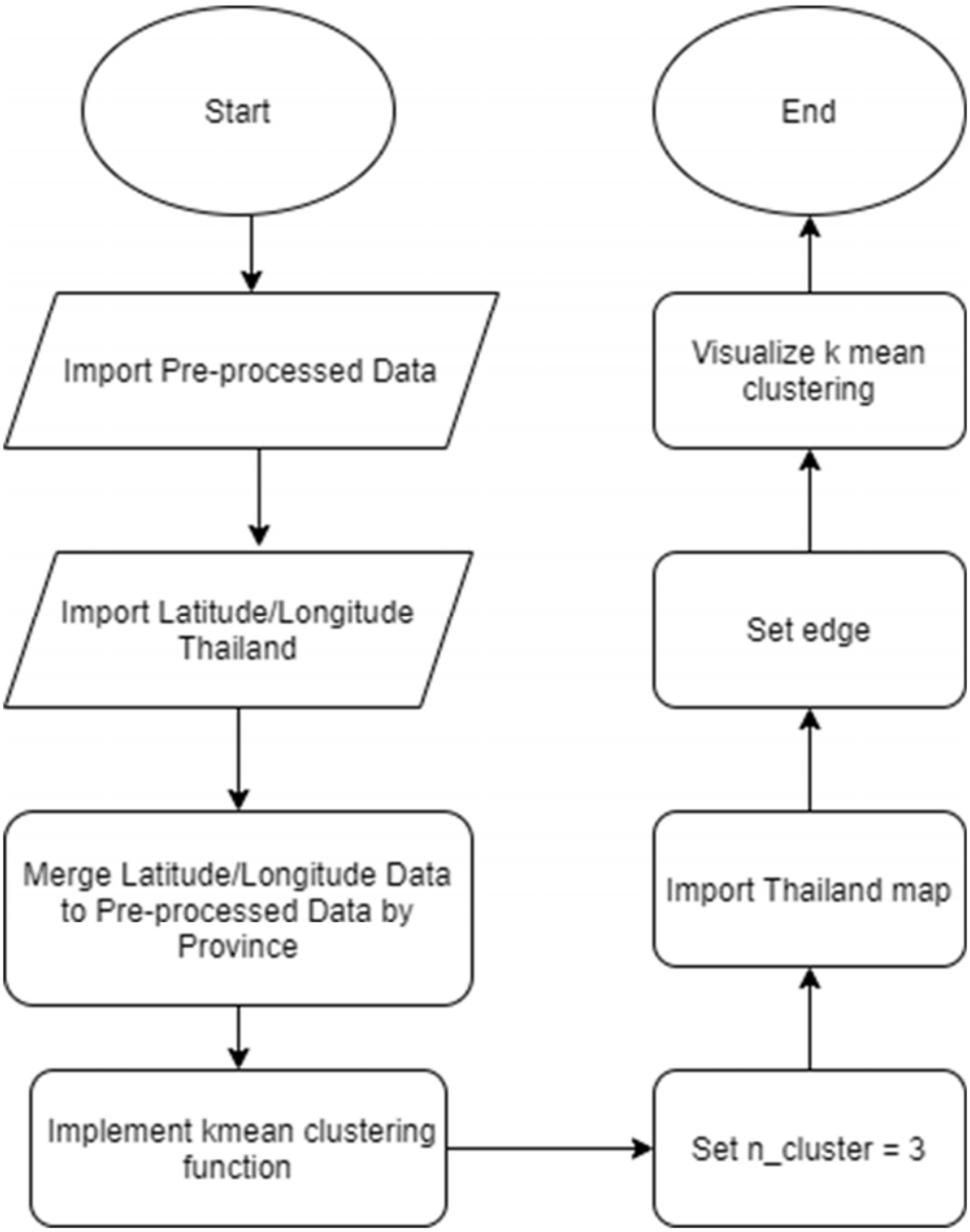


Figure 9: Flow chart diagram of create a clustering visualization model

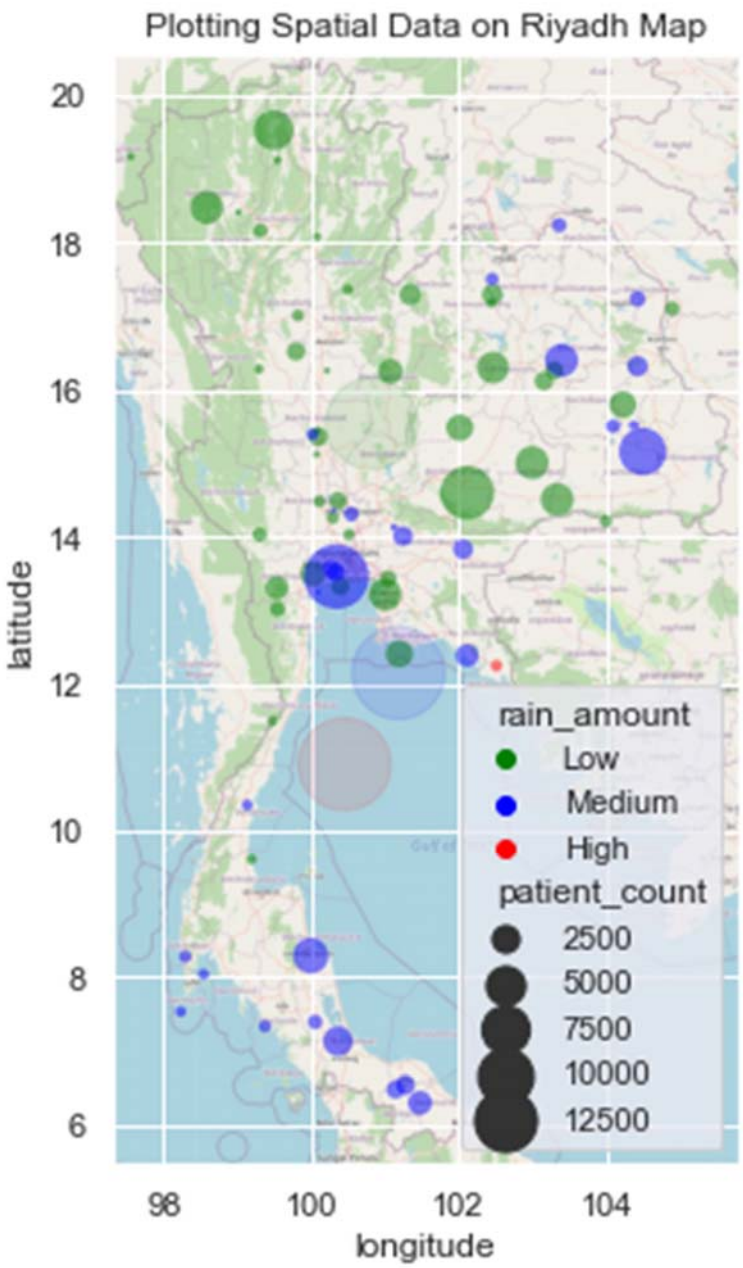


Figure 10: Map visualization of number of patient compared to amount of rain in each group

We grouped 3 clusters by latitude, longitude, rain_amout for Green color is low rain amount (1181), Blue color is medium rain amount (2155) and red color is high rain amount (3647). Size of the circle is demonstrated by the percentage of deaths by Dengue. Faded color circle means centroid.

Data visualization by heatmap plotly

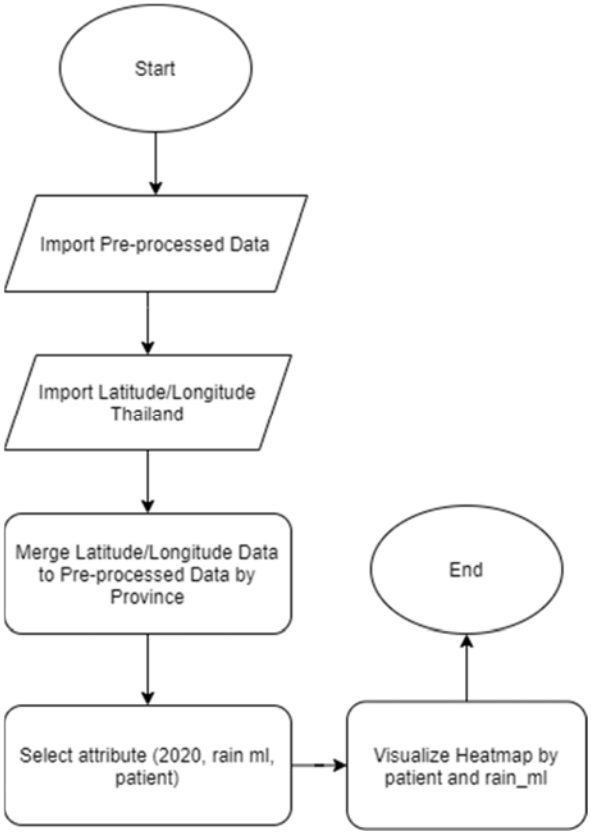


Figure 11: Flow chart diagram of create a heat map visualization model

Patient heatmap in 2019

Figure 12: Map visualization of number of patient in 2019

Rainfall heatmap in 2019

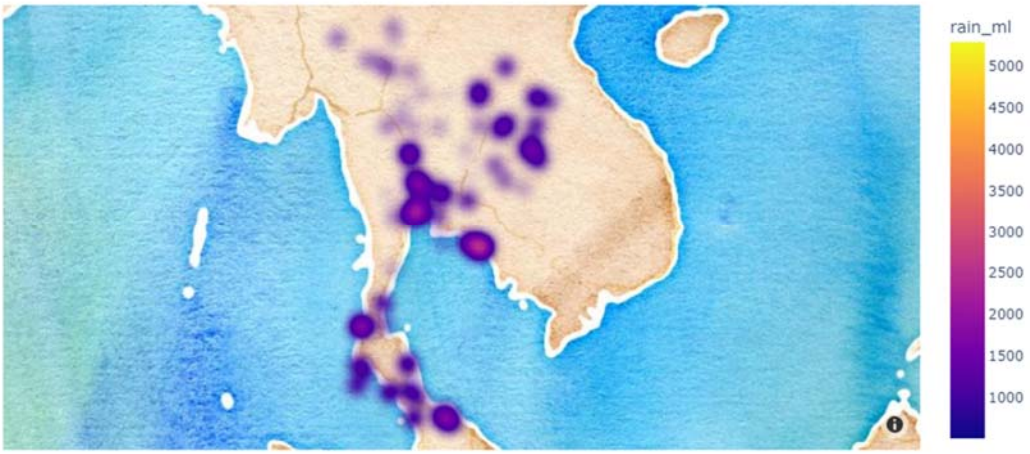


Figure 13: Map visualization of number of rainfall in 2019

From the heatmap, we can see that rainfall and number of patients are related. The concentration of patients is in Bangkok, Chiang Mai, and Korat. Rainfall dataset from the government should be updated.

5) Project Results:

Regression Model

Method	Model	R2 score (Train)	R2 Score (Test)	RMSE	MAE
1	Linear Regression (original)	57	49	812.88	622
2	Repeat K-foldsCross-Validation	57	52	975.68	671.32
3	Recursive Feature Elimination (RFE)	56	52	970.60	605.22
4	Grid Search CV	55	55	864.66	601.58
5	Polynomial Regression	83	78	753.65	544.95

For the results in this section, we used the default settings for all machine learning approaches imported from the sklearn package.

Overall, we have experimented with various machine learning approaches in predicting dengue patients for each province that have a different environmental condition. We showed that Polynomial out-performs other approaches and achieves an R-squared score greater than 0.78. In future work, we would like to improve our collecting data and add more factors that more correlate to a dependent variable for greater and reliable results in prediction.

6) Conclusion and Future Work:

Conclusion

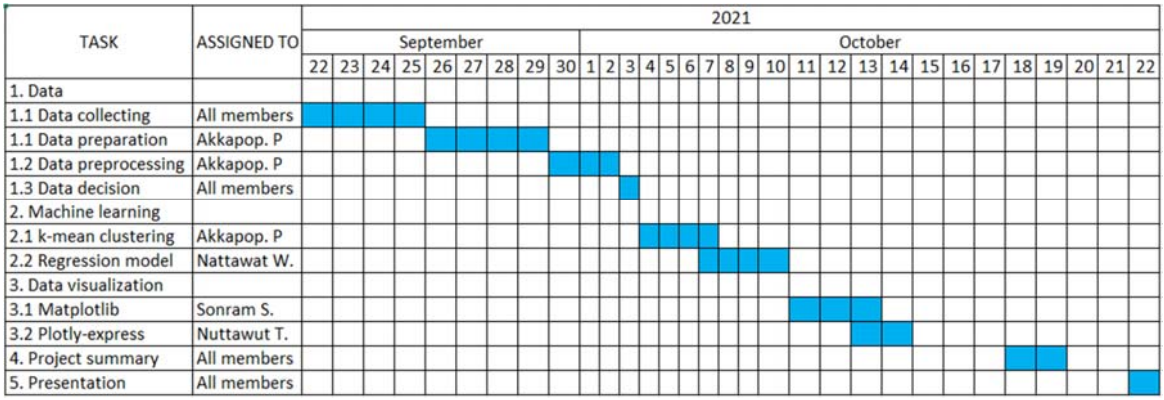
- The predicted results of the model are satisfied in some provinces, on the other hand, some provinces need more variables or more data to improve results.
- Limitation data the same kind of data is distributed. such as climate data are under the organization that does not own the data and is difficult to access information.
- For more accuracy of prediction and clustering dengue patients with adding more features like “swamp” that are a source of Aedes for growth and cause to illness.

Future work

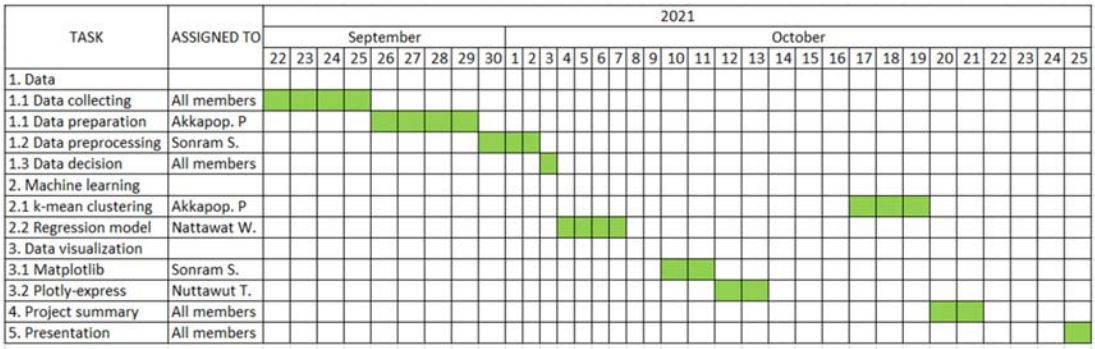
- Dengue Dashboard to follow up the situation of dengue patients to prevent or find out which areas or provinces are risks. It's open to public information for awareness and study purposes.
- Dengue Prediction Report to planning and decision-making information in dengue health policy for related organizations such as the Ministry of Public Health.

7) Project Plan VS Actual Work:

Gantt Chart (Plan)



Gantt Chart (Actual)



TASK	ASSIGNED TO	2021																														
		September									October																						
		22	23	24	25	26	27	28	29	30	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1. Data																																			
1.1 Data collecting	All members																																		
1.1 Data preparation	Akkapop. P																																		
1.2 Data preprocessing	Sonram S.																																		
1.3 Data decision	All members																																		
2. Machine learning																																			
2.1 k-mean clustering	Akkapop. P																																		
2.2 Regression model	Nattawat W.																																		
3. Data visualization																																			
3.1 Matplotlib	Sonram S.																																		
3.2 Plotly-express	Nuttawut T.																																		
4. Project summary	All members																																		
5. Presentation	All members																																		

Code

https://github.com/Phophism/python_project

8) The things that we have learned:

1. **Team Member 1 Nuttawut Thuayhanruksa**: To do this project, my role is data visualization I have learned that about data visualization from many tools such as plotly, matplotlib.Ploly is the one of new visualization. Thailand heat map is easy to understand but the heatmap's area needs latitude and longitude to plot the graph. Visualization should be reasonable and sensible to model. For that reason, visualization has many tools in the world I think I have any time to learn it.
2. **Team Member 2 Nattawat Wattanawiput**: To do this project, My role is Data modeling (Regression Model). I have learnt to help a team find different data from a data source and merge it to be a new dataset to prove the hypothesis of the project by using a data science task, building a model to predict by using Linear regression in each method to find the best model for deployment in real use.

3. Team Member 3 **Sonram Sirirat**: To do this project, my role is to collect data and collaborate to assist others in the team. First, I have learnt data science tasks as a project base and collaborated as a team to do a data task. The first challenge is to define a problem and an interesting topic. At this project we selected Dengue patients in Thailand to predict the number of patients, clustering which region has a higher chance to infect. It starts with a hypothesis that a climate can be caused to increase the patient such as rain and temperature. With data mining methods we try with different methods to compare and find which models return better results. The result is well satisfied but not the best. We realize the training data must be much more, the factor like rain does not mean the absolute answer but never say never.

4. Team Member 4 **Akkapop Prasompon**: To do this project my role is data preparation and clustering. Planning before starting to develop a software is an important process, but sometimes, the result we desired might not perform as what we expected. Thus, in the real world problem, when I create the project plan, I would add at least 50% of actual man hours into the schedule for unexpected bug fixing. However, this project has an exact due date which we can't postpone. So, technical skill is an important key for this situation. Learning and practice allow us to deal with that problem faster.

9) Reference

[1] Department of disease control (Dengue Fever). Retrieved 22 September 2021, from https://ddc.moph.go.th/disease_detail.php?d=44

[2] Lai, Y. H. (2018). The climatic factors affecting dengue fever outbreaks in southern Taiwan: an application of symbolic data analysis. Biomedical engineering online, 17(2), 1-14.

[3] Official Statistics Thailand branch of Natural Resources and Environment.Temperature, humidity, and precipitation. Retrieved 22 September 2021, from
โครงการบริหารจัดการสารสนเทศภาครัฐ เพื่อการตัดสินใจ

[4] Department of disease control (Dengue Fever dashboard) Retrieved 22 September 2021, from http://203.157.41.226/disease/Denguefever.php

[5] Open Government Data of Thailand.The fatality rate of dengue in 2017 - 2020.Retrieved 22 September 2021, from https://data.go.th/dataset/dataset-pp_36_03

[6] *Thailand latitude and longitude map*. World Map, a Map of the World with Country Names Labeled. (n.d.). https://www.mapsofworld.com/lat_long/thailand-lat-long.html.