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 640631030 Team Role: Data preparation, Visualization
2. Sonram Sirirat 640631037 Team Role: EDA, Visualization
3. Nattawat Wattanawiput 640631031 Team Role: Data collecting, Regression Model
4. Akkapop Prasompon 640631128 Team Role: Data preparation

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 discover risk groups by using dengue patients' data from the department of disease by visualizing the information into the chart using a python library such as Plotly. We expect the accuracy of our model should be higher than 60 percent.

Project Details

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 predict 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 as, we designed to visualize the age groups of patients into four specific groups including child, teenager, adult, and elder.

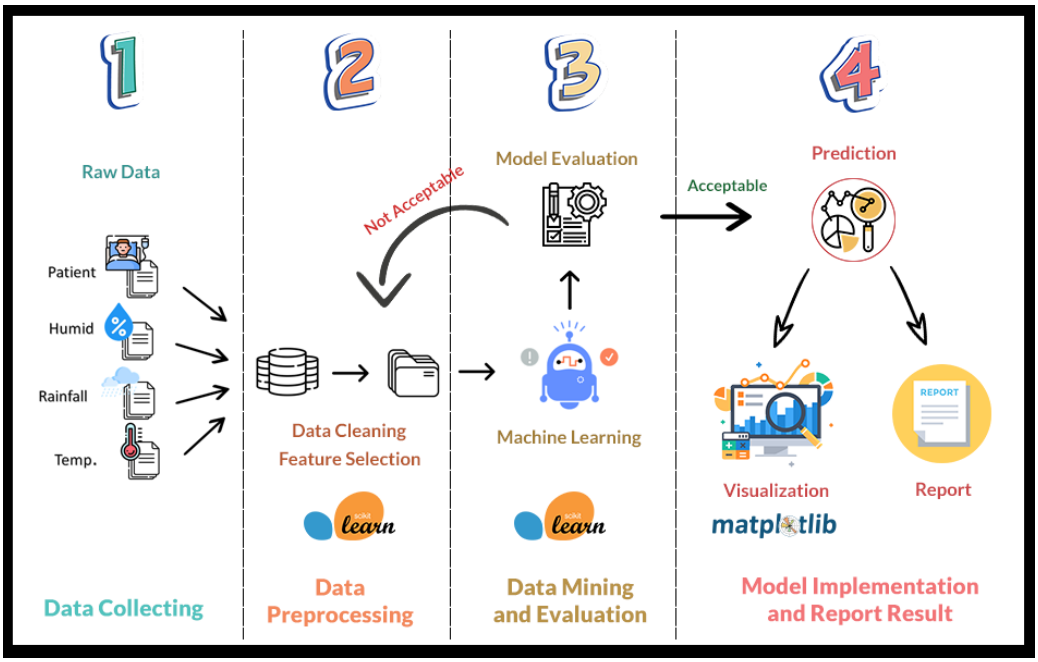

Figure 1: A diagram showing the overview of project development process

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 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 accuracy, confusion matrix, precision, recall, and f-1 score will be reported and the predicted result will be visualized using matplotlib library.

Python is the main language for developing the prediction 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]. Selectkbest 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 of the model should rate greater than 60%. We also use a data visualization technique to specify which age groups are likely to be at risk of dengue fever by using Plotly Express. The result will be represented by a scatter plot of related age group, humidity, rainfall, and temperature. As well as, an accuracy, recall, precision, and f1 score will be reported in the document.

Objectives:

1. To predict the number of dengue fever patients by average rainfall, temperature, humidity
2. To visualize information to relevant organizations by data visualization

Scopes:

1. Data Collecting
2. Data preparation
3. Data mining using linear regression model by scikit-learn library
4. Model evaluation using scikit-learn library
5. Data visualization using matplotlib library

Project Plan:

TASK	2021														
	September									October					
	22	23	24	25	26	27	28	29	30	1	2	3	4	5	6
1. Data															
1.1 Data collecting															
1.1 Data preparation															
1.2 Data preprocessing															
1.3 Data decision															
2. Machine learning															
2.1 Regression model															
3. Data visualization															
4. Project summary															
5. Presentation															

Sources

[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 https://osstat.nso.go.th/statv5/list.php?id_branch=21&Page=13

[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