

Development of Dengue Prediction Model with Neural Network

BY

Cheo Jia Jun

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION SYSTEMS (HONOURS) INFORMATION
SYSTEMS ENGINEERING

Faculty of Information and Communication Technology

(Kampar Campus)

MAY 2022

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: Development of Dengue Prediction Model With Neural Network

Academic Session: 2022

**I CHEO JIA JUN
(CAPITAL LETTER)**

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.



(Author's signature)

Address:

No.49, Jalan Mekar 8,
Taman Mekar, 86100,
Ayer Hitam, Johor

Date: 1 Sep 2022

Verified by,



(Supervisor's signature)

Ts Dr Chang Jing Jing

Supervisor's name

Date: 1 Sep 2022

Universiti Tunku Abdul Rahman			
Form Title : Sample of Submission Sheet for FYP/Dissertation/Thesis			
Form Number: FM-IAD-004	Rev No.: 0	Effective Date: 21 JUNE 2011	Page No.: of 1

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TUNKU ABDUL RAHMAN

Date: 9/9/2022

SUBMISSION OF FINAL YEAR PROJECT /DISSERTATION/THESIS

It is hereby certified that Cheo Jia jun (ID No: 18ACB05681) has completed this final year project/ dissertation/ thesis* entitled “Development of Dengue Prediction Model with Neural Network” under the supervision of Dr. Chang Jing Jing (Supervisor) from the Department of Computer and Communication, Faculty/Institute* of Information and Communication Technology, and Ts Tey Chee Chieh (Co-Supervisor)* from the Department of Digital Economy Technology, Faculty/Institute* of Information and Communication Technology.

I understand that University will upload softcopy of my final year project / dissertation/ thesis* in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.


Yours truly,

Cheo Jia Jun
(*Student Name*)

*Delete whichever not applicable

DECLARATION OF ORIGINALITY

I declare that this report entitled “Development of Dengue Prediction Model With Neural Network” is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature : 

Name : Cheo Jia Jun

Date : 9/9/2022

ACKNOWLEDGEMENTS

First of all, I would like to sincerely thank to my supervisor, Dr. Chang Jing Jing who give me a precious opportunity to engage in an Artificial Intelligence development project which is the subjects I most interest. Besides that, she also gives me a lot of valuable comment and suggestion patiently. I just want to say thank you very much for your teaching and nurturing. I also want to thank to UTAR where give me a chance to study my favorite course which is Information Technology area.

Finally, I would also like to thank my parents and friends for their constant encouragement and support, making me more confident to face all the challenges and problems of this trimester.

Abstract

The aim of this project is to develop a system that could predict the Dengue outbreak. This is done by using the prediction variables such as climate and past data. Two approaches will be used to develop the prediction models, which are Artificial Neural Network (ANN) and Generalized Additive Models (GAMs). Then, the prediction accuracy of the models will be compared. All methods can handle real life input to simulate the situation of the area where we want to predict outbreak of Dengue. It is believed that the accurate prediction of dengue outbreak can reduce the dengue case and prevent the dengue outbreak in Malaysia.

Table of Contents

Title	1
DECLARATION OF ORIGINALITY	iii
ACKNOWLEDGEMENTS	iv
Abstract	v
Table of Contents	1
<i>List of Figure</i>	3
LIST OF ABBREVIATIONS	6
Chapter 1: Introduction	7
1.1 Problem Statement and motivation	7
1.2 Project objective	8
1.3 Project Scope	9
1.4 Contribution	9
1.5 Report Organization	9
Chapter 2: Literature Review	11
2.1 Review of the Technologies	11
2.1.1 Generalized Additive Models (GAMs)	11
2.1.2 Artificial Neural Network (ANN)	12
2.1.3 Data Collection	13
2.1.4 Jupyter Notebook	17
2.1.5 Python	17
2.1.6 Budget	17
2.1.6 Libraries of Python	18
2.2 Review of the Existing System	20
2.2.1 Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia	20
2.2.2 Prediction of Dengue Outbreaks Based on Disease Surveillance and Meteorological Data	21
2.2.3 A dengue fever predicting model based on Baidu search index data and climate data in South China	21
2.2.4 Superensemble forecasts of dengue outbreaks	22
2.2.5 Application of Artificial Neural Networks for Dengue Fever Outbreak Predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico	22
2.2.6 Prediction of Dengue Cases in Paraguay Using Artificial Neural Networks	24
2.2.7 Prediction of dengue outbreak in Selangor Malaysia using Machine learning techniques	24
2.2.8 Utilizing Artificial Intelligence as a Dengue Surveillance and Prediction Tool	25

2.2.9	Artificial Intelligence Model as Predictor As Predictor For Dengue Outbreak	25
2.2.10	Prediction of Dengue Incidence in the Northeast Malaysia Based on Weather Data Using the Generalized Additive Model.....	27
2.2.11	Prediction of dengue outbreaks based on disease surveillance, meteorological and social-economic data.....	28
Chapter 3 System Model.....		29
3.1.1	Methodology	29
3.2	Formula	30
3.2.1	Formula of Artificial Neural Network.....	30
3.2.2	Formula of Linear Generalize Additive Model.....	30
3.2.3	Formula of Root Mean Squared Error	31
3.2.4	Formula of Mean Absolute Error	32
3.2.5	Formula of R-squared.....	32
Chapter 4 Implementation detailed.....		34
4.1	System Design/Overview	34
4.2	Hardware Setting up	36
4.3	Software Setting up.....	36
4.3.1	Software.....	36
4.3.2	Data Exploration	37
4.3.3	Data preprocessing.....	39
4.3.4	Artificial Neural Network and Generalize Additive Model building and Training.....	40
4.4	Fine Tune	42
4.4	Implementation issues and challenges	43
Chapter 5 Evaluation		45
5.1	Discussion of the result before fine tune	45
5.2	Evaluation	47
5.3	Comment, Highlight, Model Selection	50
Chapter 6 Conclusion		52
References.....		53
Weekly Log.....		59
Poster		65
PLAGIARISM CHECK RESULT		66
FYP2 CHECKLIST.....		73

List of Figure

<i>Figure 1 Example of GAM</i>	<i>11</i>
<i>Figure 2 Example of Artificial Neural Network</i>	<i>12</i>
<i>Figure 3 Annually Dengue Fever Trend 1999-2019</i>	<i>13</i>
<i>Figure 4 The Rainfall Data of Different State of Malaysia</i>	<i>14</i>
<i>Figure 5 The population of Different State of Malaysia</i>	<i>15</i>
<i>Figure 6 The Temperature Data of Different State in Malaysia</i>	<i>16</i>
<i>Figure 7 The Latitude and Longitude of Different State in Malaysia</i>	<i>16</i>
<i>Figure 8 Relative MAE</i>	<i>20</i>
<i>Figure 9 GAMs Formula to Predict Dengue in The Project</i>	<i>21</i>
<i>Figure 10 ANN schematic of this project</i>	<i>23</i>
<i>Figure 11 The Formula Using in the Project</i>	<i>26</i>
<i>Figure 12 Formula of GAM</i>	<i>27</i>
<i>Figure 13 Features of Data set of the Project</i>	<i>28</i>
<i>Figure 14 The Concept Map of System</i>	<i>29</i>
<i>Figure 15 Formula of Artificial Neural Network</i>	<i>30</i>
<i>Figure 16 Formula of Linear Generalize Additive Model</i>	<i>30</i>
<i>Figure 17 Formula Linear Regression</i>	<i>30</i>
<i>Figure 18 Formula of RMSE</i>	<i>31</i>
<i>Figure 19 Formula of Mean Absolute Error</i>	<i>32</i>
<i>Figure 20 Formula of R-Squared</i>	<i>32</i>
<i>Figure 21 Top-down System Design Diagrams</i>	<i>35</i>
<i>Figure 22 Evaluation of Model</i>	<i>38</i>
<i>Figure 23 The Relationship Between the Features</i>	<i>39</i>

<i>Figure 24 Visualize the relationship of Features</i>	<i>39</i>
<i>Figure 25 Visualization of Artificial Neural Network</i>	<i>40</i>
<i>Figure 26 Visualization of GAM</i>	<i>41</i>
<i>Figure 27 Fine Tune of the ANN</i>	<i>43</i>
<i>Figure 28 Simulation of Grid Search use in the GAM</i>	<i>43</i>
<i>Figure 29 Actual Outbreak of Dengue</i>	<i>45</i>
<i>Figure 30 Visualize How the Prediction to the Best Fit of Line</i>	<i>46</i>
<i>Figure 31 Prediction of GAM Before Fine Tune</i>	<i>46</i>
<i>Figure 32 Visualization the Best Fit Line of the GAM Before Fine Tune</i>	<i>47</i>
<i>Figure 33 Evaluation of Artificial Neural Network Before Fine tune</i>	<i>49</i>
<i>Figure 34 Evaluation of Linear Generalize Additive Model Before Fine Tune</i>	<i>49</i>
<i>Figure 35 Evaluation of Artificial Neural Network After Fine Tune</i>	<i>49</i>
<i>Figure 36 Evaluation of Linear Generalize Additive Model After Fine Tune</i>	<i>50</i>

Table 1 Hardware Setup

List of Table

36

LIST OF ABBREVIATIONS

<i>ANN</i>	Artificial Neural Network
<i>GAMs</i>	Generalized Additive Models
<i>RF</i>	Random Forest
<i>ReLU</i>	Rectified Linear Unit
<i>MAE</i>	Mean Absolute Error
<i>ML</i>	Machine Learning
<i>ARIMA</i>	Auto Regressive Integrated Movin Average

Chapter 1: Introduction

Since the 21st century, fighting various diseases or viruses has become an important part of human beings. The most common disease among them is dengue fever. Dengue is a mosquito-borne viral infection. The symptoms of people got infected are high fever, pain behind the eyes and in the joints, muscles, or bones, severe headache, rash, bleeding from the nose or gums and bruising easily (Dowshen, 2017.) It is transmitted by female mosquitoes mainly of the species *Aedes aegypti*. At the same time, these mosquitoes are also vectors of other disease such as chikungunya, yellow fever and Zika Viruses (WHO, 2020). Since the early 1970s, Dengue had spread to the whole Malaysia and caused a significant health burden to the population in Malaysia (Balvinder, 2017) dengue not only will affect human's health, at the same time it will affect the economy. an adjusted estimate of economic burden due to dengue illness is RM196 million per year, which is approximately RM7.14 per capital (Shepard et al., 2012). In this project, we will develop a system with high accuracy to predict Dengue outbreak which in turn help to decrease the dengue case by taking necessary precaution the results from Artificial Neural Network (ANN) and Generalized Additive Models (GAMs) will be compared. Both of the models will using the same data set that had recorded the features that can affect outbreak of dengue so it would not have the difference of input and making the prediction based on the data set, then compare the predictions that generated by both model, the highest accuracy of model will be selected as the main prediction model.

1.1 Problem Statement and motivation

Dengue fever is one of the serious diseases in Malaysia. So, dengue prevention is a major challenge facing by Malaysia government. an effective prevention of dengue in Malaysia needs the help from prediction based on past data and different condition of mosquito breeding, So, it is important to develop a system that able to predict the possibility of dengue outbreak in the Malaysia by using those data and conditions. Therefore, this project will develop neural network models based on previous data to predict the possibility of dengue outbreak in Malaysia. From many existing dengue prediction systems, it is found that climate is the most used variables because it caused

many potholes filled in the rainy season, which form a breeding ground for mosquitoes. Although the climate is the one of the most critical reason to cause dengue outbreak, but there may be others predictors variables that will cause dengue outbreak such as socioeconomic factors, temperature, weather and past data. Combining others predictor variables with the climate could make the result of prediction more accurate. So in our project, we will study the predictors variables and analysis those variables to find the other significant variables that will be affect the result of prediction.

1.2 Project objective

1) To identify prediction variables that will impact dengue outbreak.

The First objective of this project is to identify prediction variable that will impact dengue outbreak. The prediction variable will affect the accuracy of the prediction if we using the critical variable as input of ANN then we could get the higher accuracy. In other hand, if we using the wrong variable as input then we will get lower accuracy of result. So, identify the prediction variable is needed as to avoid the wrong input in ANN.

2) To develop neural network models to predict the number of dengue incidence in Malaysia

The second objective of this project is to develop a neural network model to predict the number of dengue incidence in Malaysia. The ANN and GAM that able to predict number of dengue incidence could be decrease the dengue case as we take precaution to the area where consider highest number of dengue incidence in prediction.

3) To compare the performance of ANN and GAM

The third objective is to compare the performance of ANN and GAM which can us to find out the better prediction model. This is because the better performance of prediction model can help the Malaysia Government to prepare well different suitable strategic to handler the outbreak of dengue without wasted meaningless of resources on the least effectively strategic such as implement those strategic on the

cities or states that are rarely occur outbreak of dengue, so high accuracy of prediction is very important for Malaysia Government as they just have to implement the strategic on those cities that had been predicted will occur outbreak of dengue to prevent outbreak of dengue.

1.3 Project Scope

The expected output of this project is a prediction model of dengue fever infections among Malaysians. More specifically, we will develop neural network models by comparing the accuracy of prediction result of different prediction models. This will in turn help to effectively control the mosquito breeding grounds for effective actions.

1.4 Contribution

This project is very important to Malaysia as it can help to reduce the dengue outbreak. With the help of an accurate prediction, the dengue fever infection rate could be reduced and the economic burden of dengue illness will be decreased, so the Malaysia government could have more budget to develop other economic project to improve standard living of Malaysian in turn improve the environment to decrease breeding ground of Aedes mosquito. Our Dengue prediction would be based on the existing systems because those data collected by exiting system can help this project to have more accuracy. So, by having this project we have to collect and analysis the data from exiting data butat the same time we will include other data such as geographic data, temperature, and others data that might improve the accuracy of forecast result.

1.5 Report Organization

This report is organized into 6 chapter such as, Chapter 1 Introduction, Chapter 2 Literature Review, Chapter 3 System Design, Chapter 4 Implementation Detail, Chapter 5 Model Evaluation, Chapter 6 Conclusion. The first chapter is the introduction of the project which use to show the summarize of the project to the reader to help them can easy to know the content of the project such as problem statement, project background and motivation, project scope, project objectives, project contribution, highlights of

project achievements and report organization. The second chapter is the literature review which is to carry out some of the existing project is related to this project to learn the strengths and weakness from those project. In the third chapter is to overall the system design of the project which is the programming language, model and others that use in the project. The fourth chapter is the implementation detail of the project. The Chapter 5 is the Model Evaluation which is showing how well the model working by a few of technical such as R-squared, mean squared error and others. The last chapter is the conclusion of the project which is make all the content in a nut shell and some of the discussion of the project.

Chapter 2: Literature Review

2.1 Review of the Technologies

2.1.1 Generalized Additive Models (GAMs)

The first method we use in our project is Generalized Additive Models (GAMs). This is because GAMs is a high level model system for mathematical programming and optimization. It allows us to translate the real-world optimization problem into computer code and it is flexibility as we are allowed change solvers used without model formulation change. The Figure 1 is showing the example of GAM.

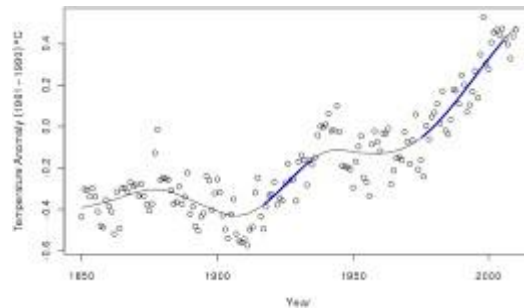


Figure 1 Example of GAMs

2.1.2 Artificial Neural Network (ANN)

Another model used in the project is Artificial Neural Network (ANN). ANN has the ability to learn and model non-linear and complex relationship as the inputs and output of most relationship in real life are non-linear and complex. The number of neurons in the hidden layer will decrease layer by layer with the shape of the inverted pyramid, and repeated attempts to determine the specific number of neurons and the missing values of the hidden layer until no further prediction is possible. After ANN learned input and their relationship, it can also infer the unseen relationship on the unseen data, so make the model can generalize and predict on the data. (Mahanta,2017) The Figure 2 is showing the Example of Artificial Neural Network.

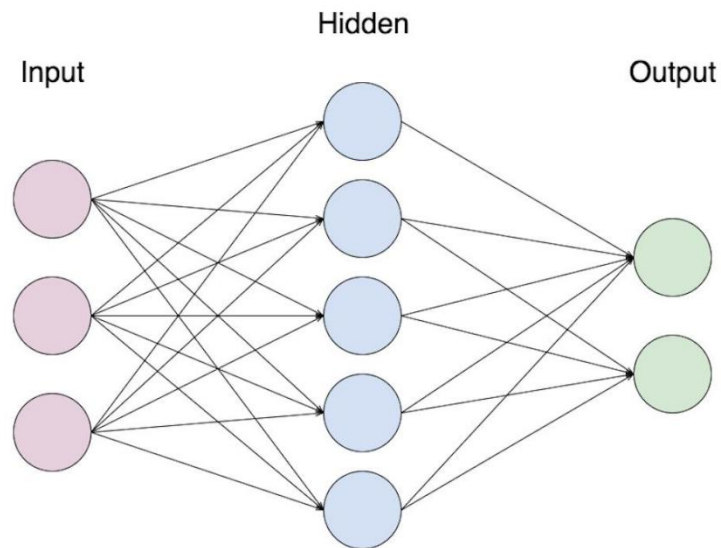


Figure 2 Example of Artificial Neural Network

2.1.3 Data Collection

Data collection is an important part of developing the model. We will use historical data as the input to the model. According to Bora (Bora, 2016), historical data could enable forecast based on billions of calculation and data point in past event to predict the event. The first data to be used is the dengue outbreak or dengue incident from the “https://data.mendeley.com/”. This website recorded the total case of dengue in different state of Malaysia which named “Dataset Denggi Malaysia.xlsx”. Besides that, the data set have recorded down the total cases of dengue in different state from 1999 until 2019. So, the data set are very import data that to help us to train the ANN as the data provide by the data set are very detail.

The screenshot shows a data table with the following structure:

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	Kemudian
NEGERI	866	866	866	866	866	866	866	866	866	866	866	866	866	866	866	866	866	866	866	866	866	866
PERAK	86	24	137	179	189	196	124	193	111	184	191	215	139	177	226	187	188	189	175	169	208	4871
SELANG	862	413	368	712	1219	1742	1862	1887	1884	1628	762	782	124	817	823	1014	1388	884	1748	1216	1249	22846
JUALAN-PAHANG	585	548	2288	945	796	2247	1617	1106	2124	2444	1834	1178	791	1013	1541	1630	2176	2481	2481	4071	4171	58216
PERAH	1047	187	1291	1628	4418	1108	1892	2887	2887	4129	1734	2288	1411	1716	2118	1725	1666	1777	1411	1734	1716	48416
SARAWAK	1974	1418	1618	1981	4519	1918	1775	1124	1181	2182	1626	1647	1718	1111	2382	1620	1118	1182	4028	4138	7241	102776
MELAYU-SELANG	1129	419	2618	4121	4017	1288	1888	1788	1188	1846	1746	426	2018	2418	2118	1181	1111	1111	1111	1111	1111	1111
NUSANTARA	427	741	941	1124	2012	1104	1170	1171	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181
NEGERA	229	114	129	111	411	861	869	817	712	818	148	145	418	448	1148	1718	1618	2188	1451	712	2118	2118
JOHOR	917	1418	1241	1912	1189	1618	1718	1718	1718	1718	1718	1718	1718	1718	1718	1718	1718	1718	1718	1718	1718	1718
PERANG	782	818	1817	1248	1887	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181
TERENGGANU	411	121	171	1176	811	1181	741	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181
KELANTAN	441	148	1887	1778	2018	1481	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181
SARAWAK	471	744	461	884	884	1411	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181
SELANG	148	111	162	162	712	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181	1181
LABUAN	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
MALAYSIA	11144	7101	14148	17167	11141	11881	11816	11816	11816	11816	11816	11816	11816	11816	11816	11816	11816	11816	11816	11816	11816	11816

Figure 3 Annually Dengue Fever Trend 1999-2019

Besides that, another data which is very important in this project is the weather data and we could get the weather data from “www.data.gov.my”, this website provides a platform to connect government agencies as data owners with various types of open data users. Increase the exposure, awareness and participation of government agencies, companies, and communities related to the implementation of the Malaysian Open Data Program, and its value in improving the provision of government services to the people. So, we could easily get the official climate data set from this website. For example, the Figure 4 had shown the rainfall data in Johor. The dataset not only provide the rainfall data in Johor, it also provides other rainfall data of other state from 2014 to 2020 and the rainfall data is one of the important factors that could be significantly affect outbreak of dengue. Besides that, this data also provides the population data of different state from 2016 to 2020. The population also is one of the important factors that could be significantly affect the dengue outbreak. The Figure 5 shows the population that record in the dataset.

2014-2020 :Daily Projected Rainfall CCSM3B1...

URL: <https://www.data.gov.my/data/dataset/5dbe1e4c-506d-4260-b471-b11594679e42/resource/58df99f7-814b-4217-a946-93b904a26678/download/daily-pro...>

Daily Projected Rainfall CCSM3B1 for 2014-2020 by State in Peninsular Malaysia

Data Explorer

Fullscreen Embed

Add Filter

_id	State	Year	Month	Day	Rainfall ...
1	Johor	2014	1	1	6.148098
2	Johor	2014	1	2	4.845438
3	Johor	2014	1	3	1.268451
4	Johor	2014	1	4	0.441532
5	Johor	2014	1	5	2.200269
6	Johor	2014	1	6	2.300741
7	Johor	2014	1	7	0.675724
8	Johor	2014	1	8	3.760135
9	Johor	2014	1	9	3.335825
10	Johor	2014	1	10	7.823872

Figure 4 The Rain fail Data of Different State of Malaysia

dataset

Muat turun

URL: <https://www.data.gov.my/data/dataset/f5a78c54-63c2-435c-b08a-d6b04f08aa8e/resource/184aee01-8562-42bc-9b05-75e601e445f0/download/m-20210226124505...>

This dataset shows the Population by state, administrative district and sex, 2016-2020. Footnote The mid-year population estimate based on the Population and Housing Census Malaysia 2010. Summation may differs due to rounding Source : DEPARTMENT OF STATISTICS MALAYSIA

Data Explorer

Fullscreen Embed

Grid Graf Peta 1480 records 1 - 100 Search data ... Go » Filter

Year	Country...	Adminis...	Sex	Populati...
2016	Malaysia		Female	15358.3
2016	Malaysia		Male	16553.9
2016	Johor	Batu pahat	Female	224.2
2016	Johor	Batu pahat	Male	238.2
2016	Johor	Johor Ba...	Female	716.2
2016	Johor	Johor Ba...	Male	820.7
2016	Johor	Kluang	Female	146.7
2016	Johor	Kluang	Male	186.4
2016	Johor	Kota tinggi	Female	103.1
2016	Johor	Kota tinggi	Male	114.2
2016	Johor	Kulai	Female	129
2016	Johor	Kulai	Male	150.5
2016	Johor	Mersing	Female	37.3
2016	Johor	Mersing	Male	42.6
2016	Johor	Muar	Female	129.9
2016	Johor	Muar	Male	144
2016	Johor	Pontian	Female	82.2
2016	Johor	Pontian	Male	90.8
2016	Johor	Segamat	Female	103.2
2016	Johor	Segamat	Male	106.8

Figure 5 The Population data of Different State of Malaysia

The temperature is one of the factors that could be significantly affect the outbreak of dengue and we can get the temperature data from an organization which is “Climate Change Knowledge Portal (CCKP)”. The organization provide an online platform from which access and analyze comprehensive data related to climate change and development. It provides all the annually temperature data of the world where include the annually temperature data of different state in Malaysia and the provided data are stored from 1901 until 2020 which is huge amount of data. The annually temperature data of different state in Malaysia is provided on their official website and we can get the data from the website. The Figure 6 is showing the temperature of different state in Malaysia.

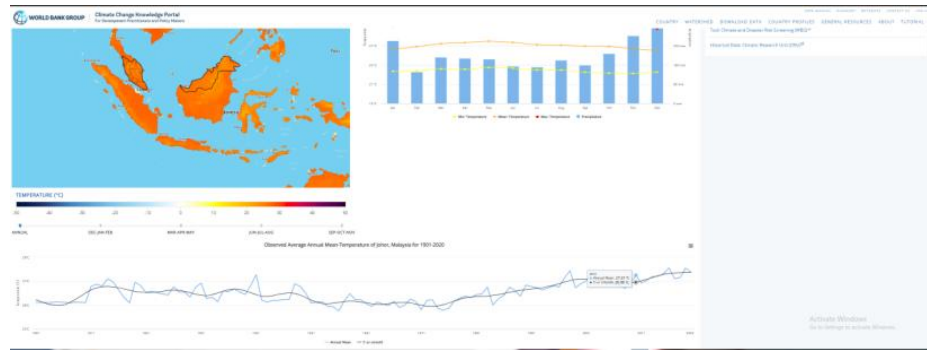


Figure 6 The Temperature Data of Different State in Malaysia

To visualize the total cases of different state in Malaysia, we need to get the latitude and longitude of each state. The “LatLong.net” could help use to get the latitude and longitude of different state in Malaysia as this website provide the latitude and longitude of any place of the world by clicked the map that provide by the “LatLong.net”, and show the latitude and longitude of clicked place. The Figure 7 is showing the latitude and longitude of different state in Malaysia.

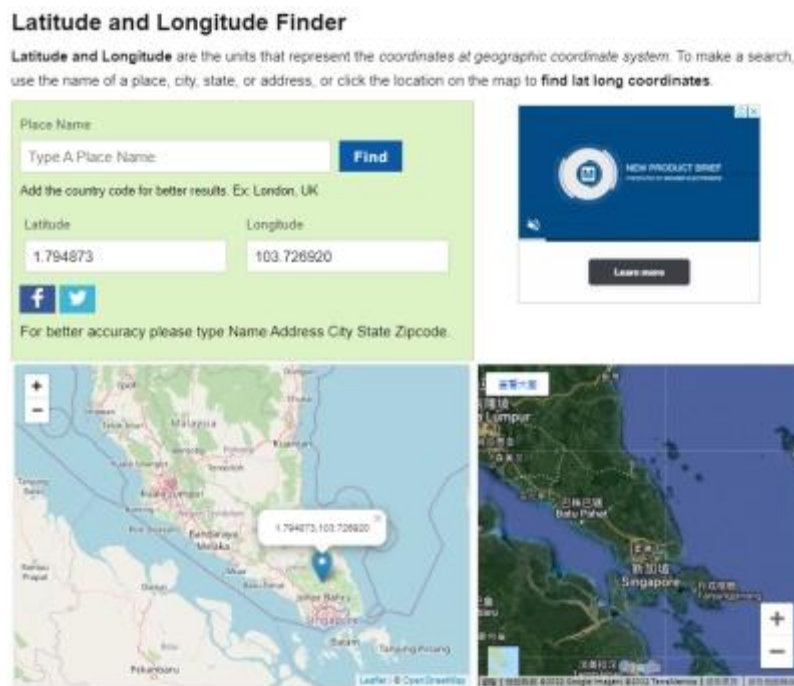


Figure 7 The Latitude and Longitude of Different State in Malaysia

2.1.4 Jupyter Notebook

The software using in the project is the Jupyter notebook as our compiler. The Jupyter Notebook is a web-based interactive development environment for notebook, code, and data. The Jupyter Notebook are very flexible as it can import different useful library in the project, so it is a very convenience tool in our artificial neural network. Besides that, it also can visualize the data and compare the result with our expectation by using the imported library.

2.1.5 Python

The programming language that involved in our project is Python. This is because Python has a great library ecosystem such as Scikit-learn for handling basic ML algorithms, Pandas for high- level data structures and analysis and others library could help us to archive the objective of this project. Another reason we choose Python to achieve our project is because the source code of Python is flexible which mean we can combine python and other languages to reach the objective and we are allowed that we could choose the programming styles which make us have a comfortable coding way such as functional style, object-oriented style and other style. The last reason we choose Python as our programming language is readability which mean it is simply understand and others developers also allowed to understand what the source code in our project.

2.1.6 Budget

In the current status of our project does not require any budget as the Python can be download free from the Python own website. At the same time, the GAMs and ANN also is free as it is one of method that can be implemented by using Python.

2.1.6 Libraries of Python

2.1.6.1 Numpy

The Numpy is one of the scientific computing of Python which able to provide a multidimensional array object, derived objects and have various routines that use to fast operation on arrays. Besides that, it also allows the user able to use mathematical, sorting, selecting, Input and Output file, random simulation and others that to help to data pre-processing and machine learning.

2.1.6.2 Pandas

The pandas also are one of the library of Python which use to handler data science, data analysis and machine learning task. Besides that, it able to handler various of time consuming, repetitive tasks that involve with the data such as data cleansing, data file, normalization, merges and joins, data visualization and other useful function.

2.1.6.3 Matplotlib

The Matplotlib is one of the library of Python that use to generate a static, animated, and interactive visualization. This library able to customize the visualization, zoom, pan, update of a interactive figures, and others. It can help our project to visualize outbreak of dengue in different state of Malaysia.

2.1.7.4 Sklearn

The sklearn is one of the library of Python that provide various of unsupervised and supervised learning algorithm. It provides many useful and convenience function that to help user to build a machine learning such as regression, classification, clustering, model selection, and preprocessing.

2.1.6.4 Keras

The Keras is the API written in Python that use to implement the neural network. It allows developer to implement the neural network easily and it could provide multiple back-end neural network computation. The advantages of Keras are able to run smoothly on both CPU and GPU, valid to most of the neural network include artificial neural network model, and modular in nature.

2.2 Review of the Existing System

2.2.1 Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia

This project was used the Random Forest(RF) to predict the mosquito-borne disease. The Random Forest were developed by two department level which are department level at national and in Colombia to predict weekly dengue cases for 12-week ahead. They were using Autoregressive Integrated Moving Average(ARIMA) as baseline (Zhao,2020) The ARIMA was used to explains a given time series based on its own past values, its own lags and the lagged forecast errors, so the equation could become the value that to predict future. (Prabhakaran, no date). After that, they compare the errors of the nationally pooled Random Forest model with Artificial Neural Network (ANN) to estimate the important change in different predictors based on predict range. The ANN is consisted of 3 level. The first level has one input layer, second level consider with 3 hidden layer, and last is consider with one output layer. This project use Rectified Linear Unit(ReLU) to solve the problem of gradient missing and had set “dropout” to avoid overfitting. The number of neurons in the hidden layer decreases layer by layer with the shape of the inverted pyramid and repeatedly try to determine the specific number of neurons and the miss value of hidden layer, until the Mean Absolute Error (MAE) cannot be forecast further. Figure 8 is relative MAE (RMAE) which to improve the intuitive interpretation and improve the prediction performance of the comparison model in different departments and prediction ranges to evaluate model accuracy. The A represented the ML model and B represented the baseline ARIMA. Although the Random forest can improve on single decision trees, but could be use more complex techniques, because the accuracy of prediction of complex problem usually worst then other model. (Hoare, no date)

$$RMAE_{A,B,h} = \frac{MAE_{A,h}}{MAE_{B,h}}$$

Figure 8 Relative MAE

2.2.2 Prediction of Dengue Outbreaks Based on Disease Surveillance and Meteorological Data

This project presents a machine learning-based methodology capable of provide predict estimates of dengue forecasting in districts of Thailand by leveraging data from different data sources. To improve the accuracy of prediction, this project used the best combination of predictors which are meteorological data, clinical data, lag variable of disease surveillance socioeconomic data and data that encoding spatial dependence on dengue transmission. The strength of this model is allows combining different predictor variables to make predictions one month in advance, and describes the statistical significance of the variables used to characterize the prediction. This project uses Generalized Additive Models (GAMs) to fit the relationships between predictors. Although it uses the GAMs, which is more effective model. but it is not sufficient in collinearity checking and out-of-sample verification shows worse results than in-sample verification. The Figure 9 had shown the GAMs formula and the element can be expressed as $\log()$ represents the natural logarithm, C represents the total number of dengue fever data, t represents the monthly time, T c represents the DTR ($^{\circ}\text{C}$), R represents the average monthly rainfall (mm), l represents the lagged variable, and ns() represents Natural cubic spline.(Jain et al., 2019)

$$\log(C_{0,t}) \sim \alpha + \sum_{l=0}^3 ns(\bar{T}_{lt}, d = 3) + \sum_{l=0}^3 ns(R_{lt}, d = 3)$$

Figure 9 GAMs Formula to Predict Dengue in this study

2.2.3 A dengue fever predicting model based on Baidu search index data and climate data in South China

The project tried to find a more effectively forecast model for estimate and predict the time of Dengue. It tries to combine the GAMs with the Autocorrelation term(AR) to propose Generalized Additive Mixed Model (GAMM). This project can help hospital managers allocate medical resources and help monitor might occurrences dengue

outbreak. By creating an ANN set and comparing the similarity of the weight results of model inputs, the neural path strength feature selection (NPSFS) abnormal events. At the same time, this project introduced the Dengue Baidu Search Index as variable such as temperature, relative humidity and precipitation. However, those data cannot represent all internet data because the Baidu search Index provides dimensionless data and the website does not give a specific calculation method, indication that these data only partially reflect the trend of dengue fever cases (Liu et al., 2019).

2.2.4 Superensemble forecasts of dengue outbreaks

This project developed 3 distinct forecasting systems for dengue outbreak in 2 places which are San Juan and Puerto Rico, and use Bayesian averaging methods to combine and create superensemble. The three forecasting system are $F1_N$ is used for predictions generated using the model inference framework, $F2_N$ is used for predictions generated using Bayesian weighting of historical outbreaks, and $F3_N$ is used for predictions derived from historical likelihood functions. They use the Bayesian averaging methods to combine the prediction from these systems and create superensemble forecasts. The strength in this project is combining these individual predictions, super ensemble predictions can be generated that can offset some individual system biases while preserving the reliable aspects of each prediction. Although this method has more accuracy for the prediction result, but due to the limited ability of the prediction system to adapt to the observation results, it is difficult to derive the true value of an outbreak similar to the long-term average. (Yamana et al., 2016)

2.2.5 Application of Artificial Neural Networks for Dengue Fever Outbreak

Predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico

In this project, the researcher had applied Artificial Neural Network trained with genetic algorithms to forecast dengue outbreak in the state of Yucatan, Mexico and in San Jun which were consider dengue outbreak. From the Figure 10, we could know the various variable use in the project are air and sea surface temperature, humidity, precipitation, previous dengue cases and size of population. The aim of the project is to understand the

dynamics of dengue and improve epidemiological surveillance and warning the area that method is used to identify the most relevant inputs However, this project also does not consider some important factors when the researcher defining the predictive power of Artificial Neural Network such as serotype, population movement, rainfall and etc. (Laureano- Rosario et al., 2018)

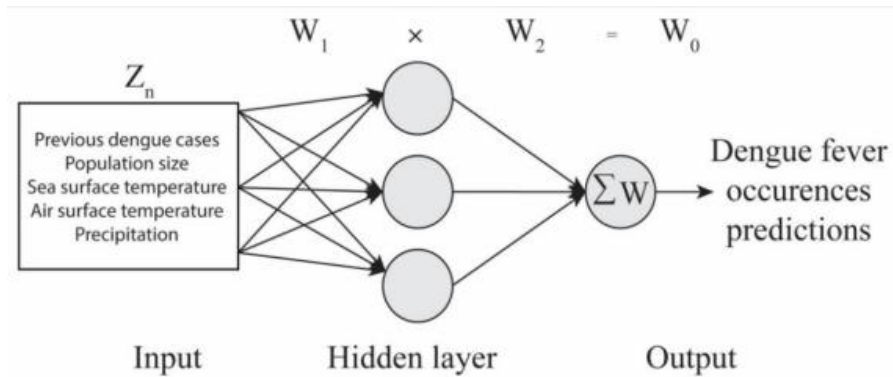


Figure 10 ANN schematic of this project

2.2.6 Prediction of Dengue Cases in Paraguay Using Artificial Neural Networks

This project had developed a predictive model based on number of dengue case and identification other covariates that affect this number. Besides, the model can predict at least one week in advance, so as to increase the preparation time of the authorities to formulate an action plan as not enough time. The predictive model is implemented by Artificial Neural Network to optimize the value of parameters necessary. The ANN have the ability to analyze the impact of climate variable related to dengue and obtain a high precision in the prediction of number of cases. However, lack of web applications that allow the authorities to predict the reasonable expected time of the number of dengue fever cases and determine the variables impact prediction. (Ughelli et al., no date)

2.2.7 Prediction of dengue outbreak in Selangor Malaysia using Machine learning techniques

This project develops multiple machine learning models to predict the outbreak of dengue in Selangor, Malaysia and find the best machine learning model to predict dengue outbreaks. The factors that can affect the outbreak of dengue use in the project are temperature, wind speed, humidity and rainfall. The project is using several data mining models in analysis node, such as Decision Tress (CART), Artificial Neural Network (MLP), SVM (Linear, Polynomial, RBF) and Bayes Network (TAN). The best model has been found in the project is SVM model which have 70% of Accuracy, 95% of specificity and 56% of precision. However, the sensitive of SVM is very high which are 62.54% compared to 14.4% for imbalanced data. The high sensitive of prediction model mean any little change in the data will cause a result that totally different to previous data, and cause the accuracy become lower. Besides that, the method of obtaining models that are different such as the Logistic regression and Naive Bayes are statistical method, but the parameter estimation and logistic function of logistic regression is use the Maximum Likelihood however the Naive Bayes is use the Bayes' Theorem to calculate posterior probabilities which mean different method of obtaining models will increase the complexity of project and cause some unexpected error which could cause the accuracy become error.

(Salim,2021)

2.2.8 Utilizing Artificial Intelligence as a Dengue Surveillance and Prediction Tool

This project utilizes the Artificial Intelligence technique to build an early warning system to make a shift from prescribed to adaptive strategic for dengue surveillance. The artificial intelligence used in the project are artificial intelligence in Medical Epidemiology(AIME) which is the decision making tool that to assists data entry, retrieval, storage, and analysis for dengue vector management. To build the machine learning and deep learning algorithm of the system, this project uses programming language such as C#, R, HTML, and JS programming language. This project also has a very high accuracy which are 81.08% as it successfully predicted 37 outbreaks and cross-validated with Penang State Health Department dengue reports where 30 outbreaks. However, the prediction period of this project are very short which only have 1 month and the result only available within 400 meters radius which mean it does not have enough time to let the officers in charge of the prediction area to prepare the strategic to avoid the outbreak of dengue which is contrary to the original intention that to early warning the outbreaks of dengue and due to the results are only available within 400 meters the officers in charge of the prediction area need spend a lot of time to analysis the data.(Sundram et al.,2019)

2.2.9 Artificial Intelligence Model as Predictor As Predictor For Dengue Outbreak

The purpose of this research is to develop an early forecasting model to predict the outbreak of dengue. The factors that used in the development of forecasting model are temperature, rainfall, date of onset and date of notification and vector indices. The machine learning that to develop forecasting model is the Bayesian Network Models which is a probabilistic graphical model for representing knowledge about an uncertain domain where each node corresponds to a random variable and each edge represents the conditional probability for the corresponding random variable. (Xin-Sheng Yang,2019) The accuracy of this project is between 81% and 92% depend on the number of hidden node which mean more hidden node, higher accuracy. The Figure 11 shows the formula that use in the

project where The Joint Probability Distribution is represented by the Bayesian network, the $N = \langle G, P \rangle$ is represent the network, the DAG whose nodes X_1, X_2, \dots, X_n represent the random variables, and the relationship between these variable are represented by edge. Although the Bayesian Network have high accuracy, but it only valid when the prior knowledge is reliable which mean too much of optimistic or pessimistic expectation on the quality of prior beliefs will distort the network and cause the result to become invalid. (Choo, 2019)

$$P_B(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P_B(X_i | \pi_i) = \prod_{i=1}^n \theta_{X_i | \pi_i}$$

Figure 11 The Formula Using in the Project

2.2.10 Prediction of Dengue Incidence in the Northeast Malaysia Based on Weather Data Using the Generalized Additive Model

This project develops a Generalized Additive Model (GAM) to predict the incidence of dengue in the northeast Malaysia with the climate data correct from the Kota Bharu Station. The features that in the data set are maximum temperature, mean temperature, rainfall, and wind speed. In this project, the researcher used one of the function of GAM which call “mgcv” to evaluate the effect of weather on dengue cases by applied cubic smoothing function and Poisson family. Figure 12 shows the formula used in the project. The s represents definition of smooth term within formula of GAM, the bs indicates the type of basis penalty smoother and cr is use to penalized cubic regression spline, one of the strength of using GAM as prediction model is can handle the nonlinear relationship without certain pattern of relationship which can handle the data set use in this project. Although this project is using GAM as prediction model, but this project omission one of the important features which is the population as without population will cause the prediction cannot show how many people is suffering from dengue and do not know whether is archive the level of dengue outbreak. (Masrani et al.,2021)

$$\begin{aligned} \text{Dengue cases} \sim & s(\text{maximum temperature, } bs = cr) \\ & + s(\text{mean temperature, } bs = cr) + s(\text{rainfall, } bs = cr) \\ & + s(\text{windspeed, } bs = cr). \end{aligned}$$

Figure 12 Formula of GAM

2.2.11 Prediction of dengue outbreaks based on disease surveillance, meteorological and social-economic data.

The propose of this project is to develop a system that able to use the factors that cause outbreak of dengue and use it to forecast the outbreak of dengue. The model that use in the project is also GAM, the factors that use in the project are related to many things such as density of infected mosquitoes, immunity of people on dengue serotypes, meteorology, human related factors such as housing type, population density and others. The project combines different predictor to forecast with a one-month lead. Although, it able to predict with a one-month lead, the data set that used as the training set of model is overfitting as the data set are very complexity. The Figure 13 is showing the factors that used in the model. (Jain et al., 2019)

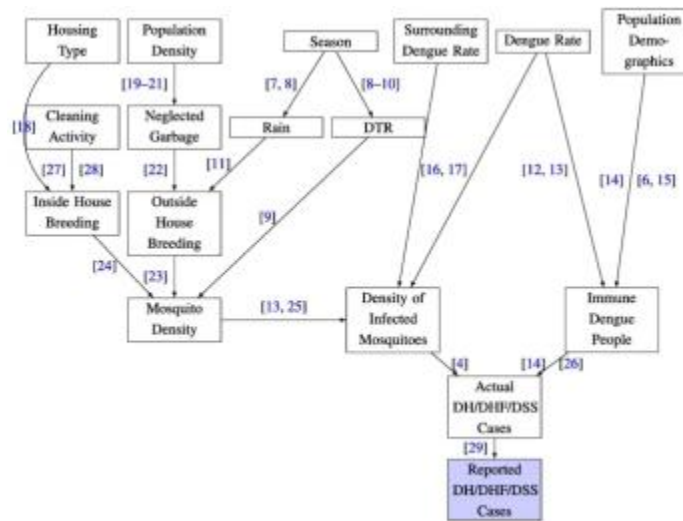


Figure 13 Features of Data Set of the Project

Chapter 3 System Model
3.1 Design specification and general work procedure
3.1.1 Methodology

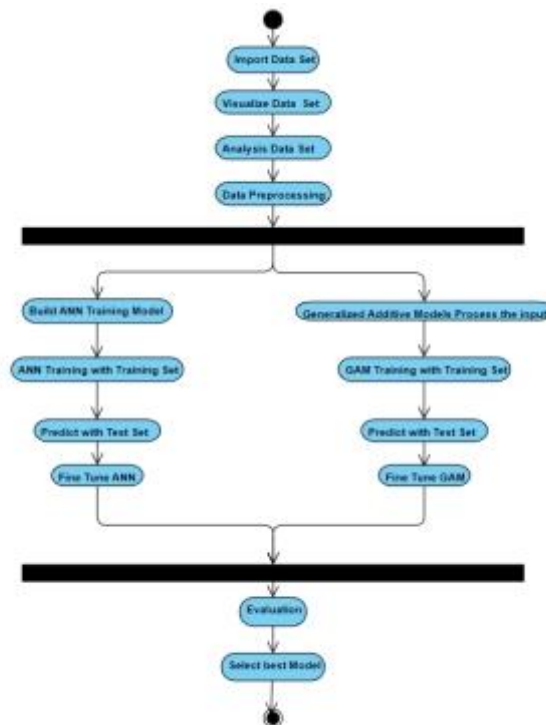


Figure 14 The concept map of system

The Figure 14 the process of the current project. The process can be divided into three phases which are data preparation, prediction model build and training, and generate result. First, we will import the data set generated by us. Then, we can visualize the data to check the quality of data and analysis the data by the different chart such as histogram chart and others. Besides that, the visualized data can help us to determine missing value of data and pre-processing the data. After preprocessing the data, we could define the artificial neural network by determine the number of hidden layer and the number of node in the hidden layer in the function which call “create_model” and use it in the “KerasRegression” to create artificial neural network. Then, train the artificial neural network by using the preprocessed data and evaluate model to determine how the data fit to model. Besides that, the another model which is GAM also will be create and step of GAM creation also is similar to the ANN which import the library then define the model and train the model with the training set. We

also will predict with the test set which use to compare with actual data set to know the accurate of prediction and evaluate the model. At the last phase, we will compare the result generate by ANN and GAM to select the best model as the main prediction model.

3.2 Formula

3.2.1 Formula of Artificial Neural Network

$$Z = \text{Bias} + W_1X_1 + W_2X_2 + \dots + W_nX_n$$

Figure 15 Formula of Artificial Neural Network

- The Z is the Artificial Neural Network
- The W is the weight of the input
- The X is the value of the input
- The Bias is the W_0 which can be known as analogous to the role of a constant in a linear function whereby the line is effectively transposed by the constant.

The steps that to perform in the artificial neural network:

1. Use the inputs and the formula of artificial neural network to generate the prediction.
2. Calculate the error term. The error term is the deviation of the actual values from the prediction values/
3. Minimize the error term

3.2.2 Formula of Linear Generalize Additive Model

$$y = \beta_0 + x_1\beta_1 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Figure 16 Formula of Linear Generalize Additive Model

$$y = \alpha + \beta x + \varepsilon$$

Figure 17 Formula of Linear Regression

The linear Generalize Additive Model is use to solve the regression problem which the formula is similar with the formula of linear regression model.

The meaning of symbol in linear regression:

- The α in the linear regression is the intercept which value of y when X=0
- The β is the slope which is amount of change in y for each until of x
- The ε is the error theme

The meaning of symbol in Linear GAM:

- The β_0 is the intercept
- The β is the slope which is amount of change in y for each until of x
- The ε is the error theme

3.2.3 Formula of Root Mean Squared Error

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\text{Predicted}_i - \text{Actual}_i)^2}{N}}$$

Figure 18 Formula of RMSE

The Figure 19 is showing the formula of the Root Mean Squared Error

The symbol meaning of Root Mean Squared Error:

- Σ = summation symbol
- The Predicted is the prediction that generated by the model
- The Actual is the data that is actual happened
- The N is the number of sample

The step to perform RMSE:

1. Get the difference between Prediction and Actual
2. Power 2 to the difference of Prediction and Actual
3. Sum them up
4. Divide by the number of error
5. Squared root to the result.

3.2.4 Formula of Mean Absolute Error

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |x_i - x|$$

Figure 19 Formula of Absolute Error

The Figure 19 is showing the Formula of Absolute Error which use to know the distance between the prediction and the best fit line.

The meaning symbol of MAE:

- The n is the number of error
- The Σ is the summation symbol
- The $|x_i - x|$ is the absolute errors.

The step to perform MAE:

1. Find the absolute error which is $|x_i - x|$
2. Sum all the absolute error
3. Divide the summation by the number of error

3.2.5 Formula of R-squared

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}$$

Figure 20 Formula of R-Squared

The Figure 20 is showing the Formula of R-Squared which use to know how fit of data on the model.

The symbol meaning of R-Squared:

- The sum squared regression is the sum of the residual squared
- The total sum of squared is sum of the distance of the data

Chapter 4 Implementation detailed

4.1 System Design/Overview

The process of the project can be divided into several phases which are data preparation, data pre-processing, develop model, model training, and use the test set to predict the result. The Figure 21 shows the top-down system design diagram of our model.

The first phase is data preparation which is extract the data from different source of data to build the data set are needed that to train the machine learning. The data needed to build the data set that to train are temperature, rainfall, population, area, total cases of dengue, latitude and longitude. Among the attributes, the temperature, rainfall, population, area and total cases of dengue are the most important attributes that can affect outbreak of dengue. Besides that, the latitude and longitude are used to visualize the data on the map.

After form the data set, the data pre-processing is needed to make sure during the training phase, the model can train properly. This is because the poorer data quality could cause many data processing efforts.

The third phase is model development which use to predict the number of dengue outbreak. The model that we use to predict the outbreak of dengue is artificial neural network. The artificial neural network is built by three layer which are input layer, hidden layer and output layer.

The fourth phase of the development is model training. In this phase, the pre-processed data will use to train the model that to predict outbreak of dengue.

The last phase is use the test set that to predict the outbreak of dengue and compare with the actual outbreak of dengue to calculate the accuracy of result, mean square error and lost.

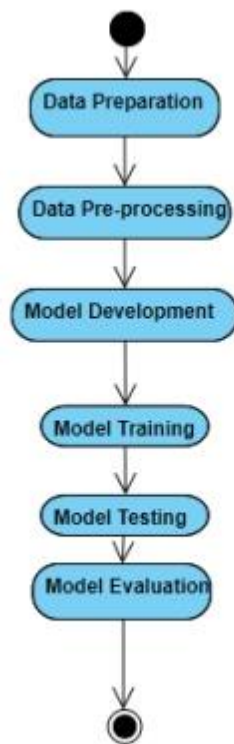


Figure 21 Top-down System Design Diagrams

4.2 Hardware Setting up

The hardware is one of the very important element in the project which a better set up of hardware can improve the process of machine learning development, the setup of the hardware that use in the project are:

Description	Specifications
Processor	INTEL CORE I7-10700F
Operating System	Window 10
Graphic Card	GIGABYTE GEFORECE RTX3060 GAMING OC 12GB DDR6
Memory	GSKILL TridentZ 2X8GB
Storage	WD 1TB Caviar Blue & SAMSUNG 970 EVO PLUS M.2 NVME 500GB SSD

Table 1 Hardware Setup

4.3 Software Setting up

4.3.1 Software

There are some of the library and software are need to download and installed before develop the artificial neural network and the generalize additive model that to predict the outbreak of dengue. The needed software is:

1. Anaconda 2021.11
2. Python 2.9.7
3. Jupyter Notebook 1.0.0
4. Dense (Library of Python)
5. Keras (Library of Pthyon)
6. Pygam Library of Pyhton)

7. Sequential (Library of Python)

4.3.2 Data Exploration

The data exploration is the process that to get the insight of data by the graphical representation of information and data by using the different graph such as pie chart, histogram and other chart. Besides that, visualize the data can help us to know what the data that is missing. So, we could conduct some data pre-processing that to handle the missing values. Besides that, visualize the data also can develop the testable hypotheses with the available data which could help to develop the artificial neural network.

The process of Data visualization:

1. Read the CSV file.
2. Visualize the data by histogram
3. Pairwise relationship of imported data set.
4. Display the data on Malaysia map
5. Implement Correlation Matrix on data set which to show the correlation coefficient of data set.

The Figure 18 is showing how serious of outbreak of dengue in different state, the red color point is representing the area is considering with very serious outbreak of dengue and we also can know the area that consider with very serious outbreak of dengue is Selangor. On other hand, the blue color is representing that the area is consider with very low outbreak of dengue and the figure show us most of the state of Malaysia is consider with low cases of dengue.

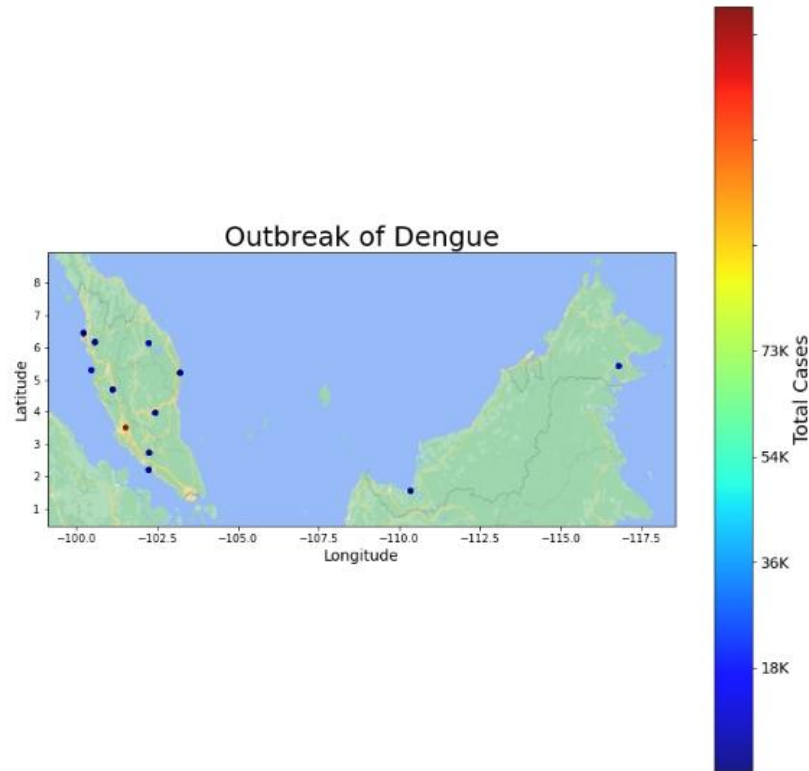


Figure 22 Total Cases plot on Malaysia Map

The Figure 24 is showing the relationship between the features, it can help us to find out the covariance in our data. When the relationship between the two features which mean if one of the features in increasing, the another features will also increase at the same time. However, if the relationship between the two features are negative is mean that when one the feature is increasing, the another features will decrease at the same time. The figure 20 is showing the visualization of relationship of features. The graph of left side is the relationship between the “Area” and “rainfall” which are positive relationship. In other word if the “Area” is increasing the “rainfall” also will increase at the same time. However, the middle graph which relationship between “temperature” and “rainfall” is a negative relationship, so when the rainfall is increasing the temperature will be decreased but this is logical as rainier days mean less sunlight and cause lower temperature. The right side of graph is showing the relationship between “Area” and “Population” and the relationship of both of the features is positive so, it means that when the “Area” is increase, the “Population” also will increase.

	rainfall	Area	population('000)	Temperature	longitude	latitude
rainfall	1.000000	0.454626	0.061675	-0.666065	-0.431289	0.212803
Area	0.454626	1.000000	0.269720	-0.505519	-0.801811	-0.259728
population('000)	0.061675	0.269720	1.000000	-0.232519	-0.306158	-0.050389
Temperature	-0.666065	-0.505519	-0.232519	1.000000	0.441148	-0.032247
longitude	-0.431289	-0.801811	-0.306158	0.441148	1.000000	0.107106
latitude	0.212803	-0.259728	-0.050389	-0.032247	0.107106	1.000000

Figure 23 The relationship between the Features

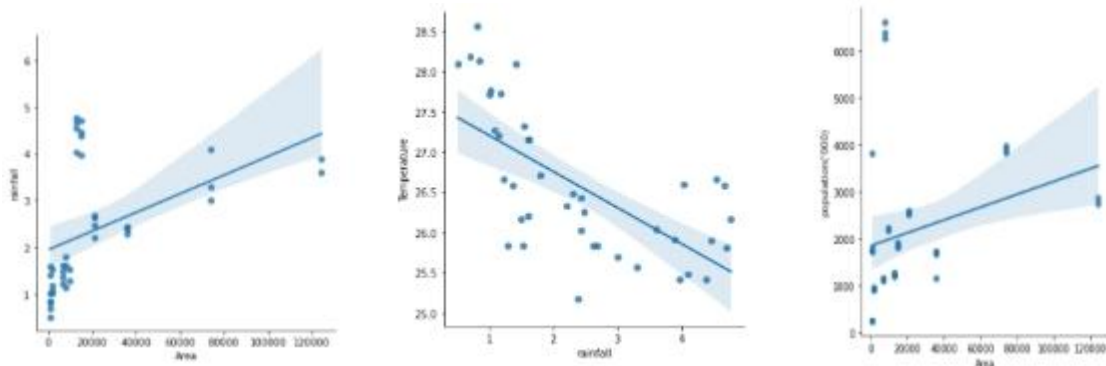


Figure 24 Visualize the relationship of Features

4.3.3 Data preprocessing

The data preprocessing is the process that to transform raw data into meaningful data for artificial neural network training as a data that does not have preprocessing, will cause the result of artificial neural network become meaningless and less accuracy. In our data set, some of the data value are missing, and the action that to handling the missing values in the data set is eliminate the objects. Then, the data set will randomly split into 80% of train set and 20% of test set. The training set and testing set conduct the Min-Max Scaler to the data set to re-scales to predetermined range between 0 and 1, and this would not change the center of distribution.

The process of data preprocessing:

1. Handler missing value by remove the objects.
2. Split the data to training set and test set

3. Normalization both of the test data and training data by MinMax Scaler

4.3.4 Artificial Neural Network and Generalize Additive Model building and Training

The model that we use in the project is the Artificial Neural Network. The ANN has 5 hidden layers where the first hidden layer has 35 nodes, second hidden layer with 30 nodes, third hidden layer with 25 nodes, fourth hidden layer with 20 nodes and the last hidden layer have 15 nodes. The output layer only has 1 node. The ANN also will be trained multiple times to determine the weights and biases that reduces the loss of data before conduct the prediction. The Artificial Neural Network is trained by using the training set data which is 80% data to entire of data set. Figure 25 shows the visualization of ANN of our project by using an online tool which is “<http://alexlenail.me/>”.

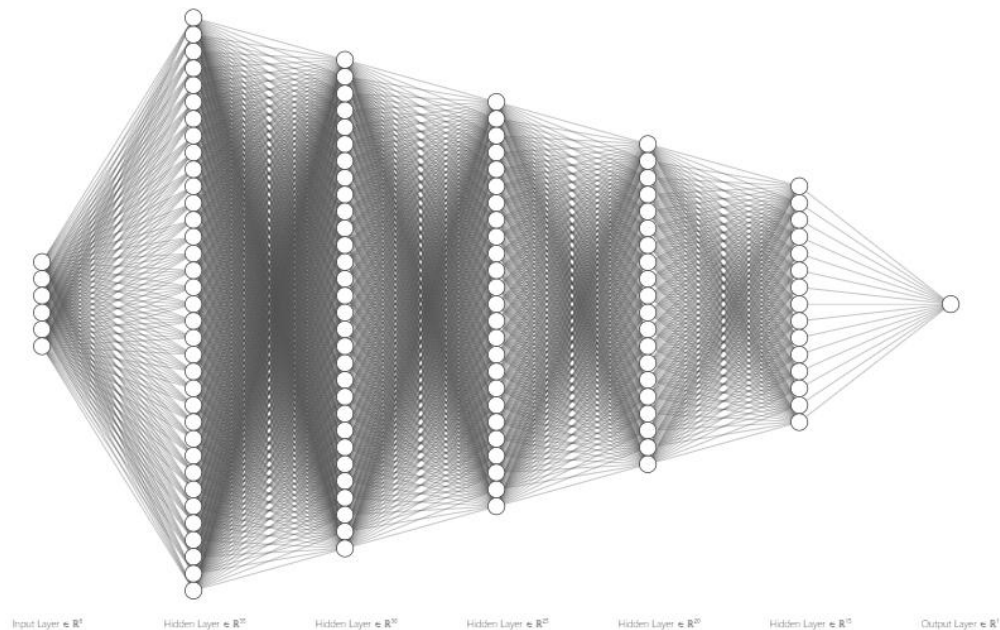


Figure 25 Visualization of Artificial Neural Network of Project

After build the Artificial Neural Network, we will build the another model that use to make the comparison with the artificial neural network which is Linear Generalize Additive Model that can be handler non-linear features. The parameter of the Linear GAM

is using the default value which are lambda is 0.6, when the lambda is higher, the line will be less wiggly. Each of the features is arranged into spline which is a special function defined piecewise by polynomials, the splines can help to generate similar result and avoid the Runge's phenomenon for higher degrees which is the problem of oscillation at the edges of an interval and it usually happen when using the polynomial. The Figure 25 is showing the visualization of GAM.

```
In [133]: from pygam import LinearGAM, s, f
gam = LinearGAM().fit(x_train, y_train)

In [134]: gam.summary()
```

```
LinearGAM
-----
Distribution:          NormalDist Effective DoF:          8.2451
Link Function:        IdentityLink Log Likelihood:       -804.0035
Number of Samples:    41 AIC:                          1628.4973
                                                             AICc:             1632.6568
                                                             GCV:              202788031.2333
                                                             Scale:            131825687.872
                                                             Pseudo R-Squared: 0.501
-----
Feature Function      Lambda      Rank      EDof      P > x      Sig. Code
-----
s(0)                  [0.6]       20        20        1.75e-05   ***
s(1)                  [0.6]       20        20        1.11e-16   ***
s(2)                  [0.6]       20        20        1.11e-16   ***
s(2)                  [0.6]       20        20        5.46e-08   ***
intercept             1           1         1         1.25e-02   *
```

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

WARNING: Fitting splines and a linear function to a feature introduces a model identifiability problem which can cause p-values to appear significant when they are not.

WARNING: p-values calculated in this manner behave correctly for un-penalized models or models with known smoothing parameters, but when smoothing parameters have been estimated, the p-values are typically lower than they should be, meaning that the tests reject the null too readily.

C:\Users\J\AppData\Local\Temp\ipykernel_26708\3358381670.py:1: UserWarning: BROWN BUG: p-values computed in this summary are 1 likely much smaller than they should be.

Please do not make inferences based on these values!

Collaborate on a solution, and stay up to date at: github.com/drvah/pyGAM/issues/183

```
gam.summary()
```

Figure 26 Visualization of GAM

4.3.5 Testing setup

The size testing set that will be using in the model for the evaluation purpose is 20% of the entire data. The 20% of the data is randomly selected as we are using the “train_test_split” function to split the training set and testing set. Since, to train the more accurate model the 80% of the data is indeed, and 20% of the data is enough to perform of the evaluation process as testing set does not used to trained model. Although, does not have any rules that mention the testing set should only 20% but if the testing set holding too much of data and cause the model does not have enough of data to train

4.4 Fine Tune

The model fine-tuning is the technique that take a model that has been trained for one given task and perform the tuning or tweaks the model and execute similar task in other word the model that are fine-tuned usually could have a better performance than previous same model. The fine-tuning method that use in this project is GridSearch CV which is the technique that find out the best combination of parameter that will be used in the model and generate more accurate result than previous model

The parameter of the Artificial Neural Network will be fine-tuning with parameter such as the `batch_size` with 5 values which are 20,40,60,80 and 80, and epochs in the range between 1 and 5000. The best parameter that get after fine tuning is `batch_size` is 60 and the epochs is 1101. After using the best parameters that are get from the Grid Search, we can find the prediction is more close to best fit line compare to previous Artificial Neural Network. The figure is showing the best combination of parameters of Artificial Neural Network.

Besides that, we also will fine tune generalize additive model with using the GridSearchCV. The GAM model has a function which call the “`LinearGam.gridsearch()`”, the “`linearGam.gridsearch()`” is a lazy and this function will not remove useless combination from the search space. However, in this project we will build a simple function that similar to the GridSearchCV. By studied the GridSearchCV, we can know that this technique is try different combination of parameter on the model and remove the useless of combination of parameter. So, based on the information we learn from the GridSearchCV, we can use the nested loop to simulate the GridSearchCV to get the best parameter of the linear GAM. The figure is showing the nested loop that can simulate the function that similar to GridSearchCV. Firstly, we also will predetermine the parameter that will be used in the model fine tune process. The first loop is the “`splines`” and the second loop is “`lam`”. The first loop will be waiting the second loop finished looping. After that, we will define the gam inside the nested loop and the parameter of GAM will be using the value of first loop and second loop. Then, train the model and calculate the `r2` of GAM and store the parameter and the `r2` score in “`result2`”. The “`result2`” will store the `r2` score to find the index of highest result. The index will use in the first array which is

“result” that storing all of the combination of parameter. The best combination of parameter that search from the nested loop is the “splines” are 14 and “lam” is 0.1. The Figure 26 and 27 is showing the fine tune of ANN and GAM

```

from sklearn.model_selection import GridSearchCV

epochs = np.arange(1, 5000, 50)
param_grid = dict(batch_size=batch_size, epochs=epochs)
grid = GridSearchCV(model, param_grid, cv=3)
grid.fit(x_train, y_train)
grid.best_params_

{'batch_size': 60, 'epochs': 1101}

```

Figure 27 Fine Tune of ANN

```

lam = [0.1, 0.01]
result = []
result2 = []
for splines in range(7, 20):
    for x, y in enumerate(lam):
        gam = LinearGAM(s(0, n_splines=splines) + s(1, n_splines=splines) + s(2, n_splines=splines) + s(3, n_splines=splines), lam=y)
        gam.fit(x_train, y_train)
        gam_pred = gam.predict(x_test)
        test_predictions = pd.Series(gam_pred.reshape(11,))
        pred_gam = pd.DataFrame(y_test, columns=["Test True Y"])
        pred_gam = pd.concat([pred_gam, test_predictions], axis=1)
        pred_gam.columns = ["Test True Y", "Model Predictions"]
        r2 = r2_score(pred_gam["Test True Y"], pred_gam["Model Predictions"])
        s = 2 * r2
        result.append((splines, y))
        result2.append(s)

bestresult = result2.index(max(result2))
print("Best parameter of GAM: ", result[bestresult])

Best parameter of GAM: [14, 0.1]

```

Figure 28 Simulation of Grid Search use in the GAM

4.4 Implementation issues and challenges

The data are the most important element that are used to build artificial neural network. However, it does not have a data set that recorded the factors that can be affect the outbreak of dengue. To build the data set have all the factors that can affect the outbreak of dengue, we have to brows website that have the related factors as more as possible to build our data set. Besides that, we have to verify the founded data whether the data are real or not and make sure the data are not outdated. During the finding data source and verify the data, it is the process that are very time consuming. Besides that, we also have a critical issue in our data set which lack of record in the data set which only have 53

records. The effect of lack of data might cause overfitting which are the model will learn as a concept from noise or random fluctuation in the training data, so it not applies to new data and negatively impact the generalization ability of the model.

Chapter 5 Evaluation

5.1 Discussion of the result before fine tune

The prediction is the output of the artificial neural network which is trained based on our data set to generate prediction of outbreak of dengue by using the testing set. Figure 28 shows the actual outbreak of dengue and prediction of outbreak of dengue. From the Figure 19, we know that most of the model prediction is very close to the actual total cases of y_{test} which mean the accuracy of the model are likely between 89% to 64%. Figure 29 also can prove that most of the model prediction and actual outbreak of dengue are very close to each other and also show that the prediction is very close to the regression best line. Besides that, to check the reliable of the prediction, we also perform some method to prove it which are mean absolute error, Root Mean Squared Error and R-Squared.

	Test True Y	Model Predictions
0	45349	56342.039062
1	823	2900.863037
2	6071	2757.593750
3	10873	7526.215820
4	288	94.620430
5	10641	7079.427246
6	2190	3113.995117
7	994	2678.360596
8	952	2793.365234
9	7932	6540.008301
10	3668	5651.188965

Figure 29 Actual Outbreak of Dengue and Prediction

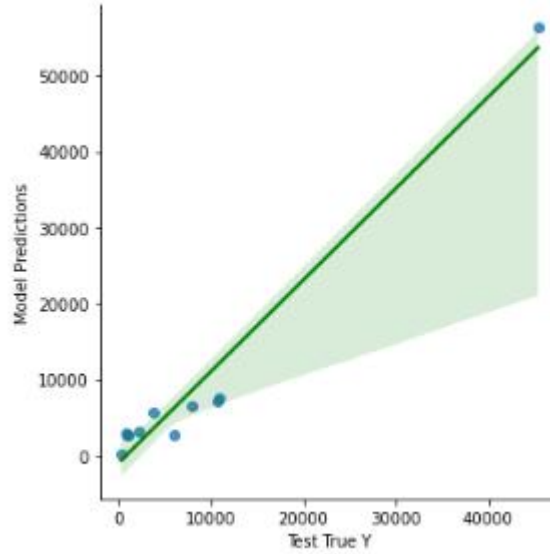


Figure 30 Visualize How the Prediction to the Best Fit of Line

On other hand, the Linear GAM that using the default value as the parameter are not perform well as from the prediction result we found that the difference between the predictions and actual output are very big which around 47% and 10% which mean the default value is not suitable to solve the problem well. From the Figure 31 we can found that, most of the prediction point is not close to the best fit line, so it can show that the GAM with the default setting is not able to handle the data. The GAM also will use the same elevation way with Artificial Neural Network to measure the performance.

	Test True Y	Model Predictions
0	45349	23796.221637
1	823	478.978333
2	6071	7596.658125
3	10873	8054.464376
4	288	2698.998057
5	10841	5984.494168
6	2190	9584.496203
7	994	11746.170359
8	952	452.300401
9	7932	8156.648299
10	3668	2843.035517

Figure 31 Prediction of GAM Before Fine Tune

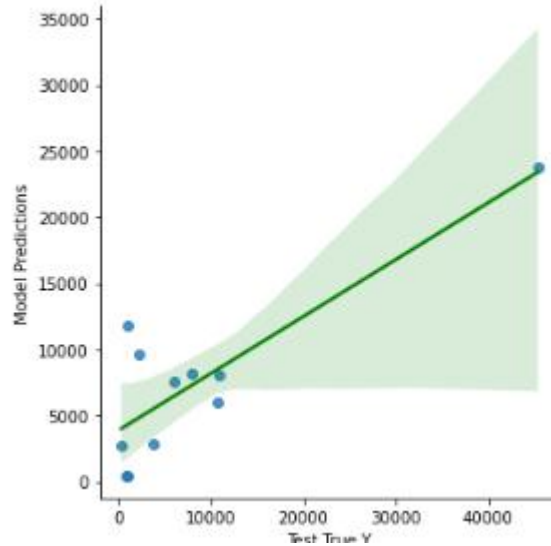


Figure 32 Visualization the Best Fit Line of the GAM before Fine Tune

5.2 Evaluation

The Figure 21 is showing the evaluation of Artificial Neural Network before fine tune. The Mean Absolute Error is used to know the absolute difference between the actual and the prediction, and the value of Mean Absolute Error in the project is 2846.4497. The Root Mean Squared Error is tool that can be used to calculate how far distance of the prediction with the line of best fit and the value of RMSE of this project is 3965.9733. The R-Squared it the statistical measure representing the proportion of variance in the dependent variable explained by one or more independent variable in the regression model and this project gain very high score of R-Squared which are around 89.675% and it mean that it quite fit to the regression. In other word, our model could make a very high accuracy prediction about the outbreak of dengue as it is very fit to the observed data.

However, the Figure 32 is showing the evaluation of Linear Generalize Additive Model before tune. From the image, we know that the Mean Absolute Error, Root Mean Squared Error and R-Squared of Linear GAM is worse than the Artificial Neural Network which are 4820.3161, 7826.7720 and 59.7917%. Various signs show that the Linear GAM with default value cannot handler well with the data, so the prediction that generate in this situation is unreliable.

Since the objective of this project is want to find the best performance between two model so we have to perform the fine to both of model to make these two model can work in the best state of both model. After use the best combination of parameter from the GridSearchCV and we get a very interesting result. The Mean Absolute Error of Artificial Neural Network is drop from 2846.4497to 2523.3070 which mean the prediction and actual are more similar to each other. The Root Mean Square also improved from 3965.9733to 3372.3942, and it's mean the distance of the prediction with the line of best fit are more close to each other. The R-squared also increase around 3% and the prediction of the Artificial Neural Network will be more accurate. The Figure 32 is showing the MAE, RMSE, and R-squared before ANN using the best parameter combination, and The figure 34 is showing the MAE, RMSE, and R-squared after ANN using the best parameter combination.

The most dramatic changing after fine tune of model is the Linear GAM which the Mean Absolute Error, Root Mean Square Error and R-squared is changes significantly after using the best combination of parameter in Linear GAM. The Mean Absolute Error is decrease 3115.8355 which is 1704.4805 and it consider around 67.65% so it's show the prediction and the actual result is quite similar. The Root Mean Squared Error is change to 2043.4267 which mean that distance of prediction and best fit line is more close to each other than the Artificial Neural Network. The R-squared of Linear GAM also higher than Artificial Neural Network which 97.2592% which it can predict the most accurate prediction. The Figure 33 is showing the MAE, RMSE, and R-squared before GAM using the best parameter combination, and The figure 33 is showing the MAE, RMSE, and R-squared after GAM using the best parameter combination.

```

In [37]: from sklearn.metrics import mean_absolute_error, mean_squared_error
mean_absolute_error(pred_df["Test True Y"], pred_df["Model Predictions"])

Out[37]: 2846.4496785944157

In [38]: mean_squared_error(pred_df["Test True Y"], pred_df["Model Predictions"])

Out[38]: 15728944.270302363

In [39]: np.sqrt(mean_squared_error(pred_df["Test True Y"], pred_df["Model Predictions"]))

Out[39]: 3965.9733068065857

In [40]: from sklearn.metrics import r2_score
r2=r2_score(pred_df["Test True Y"], pred_df["Model Predictions"])
print(r2*100, "%")

89.6759621959001 %

```

Figure 33 Evaluation of Artificial Neural Network Before Fine Tune

```

In [46]: mean_absolute_error(pred_gam["Test True Y"], pred_gam["Model Predictions"])

Out[46]: 4820.316055603066

In [47]: mean_squared_error(pred_gam["Test True Y"], pred_gam["Model Predictions"])

Out[47]: 61258361.03813569

In [48]: np.sqrt(mean_squared_error(pred_gam["Test True Y"], pred_gam["Model Predictions"]))

Out[48]: 7826.77207015355

In [49]: r2=r2_score(pred_gam["Test True Y"], pred_gam["Model Predictions"])
print(r2*100, "%")

59.791730181853076 %

```

Figure 34 Evaluation of Linear Generalize Additive Model Before Fine Tune

```

In [68]: mean_absolute_error(pred_df["Test True Y"], pred_df["Model Predictions"])

Out[68]: 2523.307004061612

In [69]: mean_squared_error(pred_df["Test True Y"], pred_df["Model Predictions"])

Out[69]: 11373043.195073368

In [70]: np.sqrt(mean_squared_error(pred_df["Test True Y"], pred_df["Model Predictions"]))

Out[70]: 3372.3942822679214

In [71]: mean_absolute_error(pred_df["Test True Y"], pred_df["Model Predictions"])
mean_squared_error(pred_df["Test True Y"], pred_df["Model Predictions"])
np.sqrt(mean_squared_error(pred_df["Test True Y"], pred_df["Model Predictions"]))
r2=r2_score(pred_df["Test True Y"], pred_df["Model Predictions"])
print(r2*100, "%")

92.53505347365939 %

```

Figure 35 Evaluation of Artificial Neural Network After Fine Tune

```

In [76]: mean_absolute_error(pred_gam["Test True Y"], pred_gam["Model Predictions"])
Out[76]: 1704.4805793749836

In [77]: mean_squared_error(pred_gam["Test True Y"], pred_gam["Model Predictions"])
Out[77]: 4175592.751933908

In [78]: np.sqrt(mean_squared_error(pred_gam["Test True Y"], pred_gam["Model Predictions"]))
Out[78]: 2043.4267180238953

In [79]: r2=r2_score(pred_gam["Test True Y"], pred_gam["Model Predictions"])
print(r2*100, "%")
97.25925804779632 %

```

Figure 36 Evaluation of Linear Generalize Additive Model After Fine Tune

5.3 Comment, Highlight, Model Selection

From the result, we know that, the proposed method is feasible. Firstly, visualize the data help us to find out the problem of the data effectively and understand the relationship between the data by visualize the data in different diagram such as the histogram and visualize the total cases on the Malaysia Map to know what state has most serious problem of outbreak of dengue. After visualizing the data, we perform data pre-processing to make our data are meaningful to the prediction model. In this phase, we remove the data that have missing value and split the data to training set and test set. Then, implement the MinMax-Scaler to the training set and test set which is the nominalize the data to make sure the data are in range between 1 and 0.

After that, we build our prediction models which are Artificial Neural Network in regression and Linear GAM with default hyperparameter to predict outbreak of dengue. We will use the pre-processing data on both of the models to test the performance. From the result, we know that Artificial Neural Network is work better than Linear GAM when both of the model is working with the default value. However, when we perform the fine tune model which is the process that to adjust hyperparameter of model to let the model can work in best performance, we found that the performance of Artificial Neural Network after used the best combination of parameter is worse than the performance of Linear GAM after using the best combination of parameter.

So, the Linear GAM model is the best model as this model also successfully generated the prediction that very close to the actual total cases of dengue. Besides that, the

prediction is also close to the best fit line which mean Linear GAM is very reliable. The three of the evaluation which are RMSE, MAE, and R-squared also show that the Linear GAM performance is better than ANN as Linear has lower error and high R-squared. Besides that, GAM not just the performance is better than the ANN but it has other advantage which is able to handle the non-linear and non-monotonic relationships between response and features. The ability to handle the non-linear complex data is mean that it does not need the polynomial to handle the complex data to make the best fit line fit to the complex data, so it can avoid the overfitting. The reason that polynomials will cause the overfitting is because it overflows easily. For example, before using the polynomial on one feature, the sample value just 1000 but after using on the feature the feature might become 1,000,000 .and it will cause the feature scaling become important as to make those data in a comparable range of value. Since it can handle the huge data that might cause the non-linear result without using the polynomial then the other developer or user who want to use the model of this project, then the just need to find a data set which have same feature that we use on the GAM. This model can help Malaysia Government to decrease the cases of the dengue as the accuracy of GAM is very high, so government can provide the strategic effectively without using too much of resources on meaningless strategic.

Chapter 6 Conclusion

In conclusion, outbreak of dengue is one of the serious problems in Malaysia and it also causes a lot of economy loss of Malaysia. This project aims to build a prediction model to predict the outbreak of dengue in Malaysia. Know in advance about the outbreak of dengue could help Malaysia Government to come up with a strategy to avoid or minimize the cause of outbreak of dengue. This can help Malaysia Government save a lot of money by effectively schedule their Dengue preventive measures and spend the money that on other policy or strategy to improve life quality of public of Malaysia. However, we know a theorem that learn in the machine learning which No Free Lunch Theorem, and it's mean does not have any model that can be handler all solution. So, this project will train two model to find out the best model which are Linear Regression Model and Linear Generalize Additive Model.

To build the both of the model, there are several phases that need to be implemented, which are data preparation, data visualization, data pre-processing, build artificial neural network, and train artificial neural network to predict outbreak of dengue. The data is a very important element that use to build artificial neural network as a model training cannot without the data and we had collected the source data from different website which are annually average rainfall, annually average temperature, area of state, longitude, latitude and the total cases of dengue of each state. After preparation of data, then we can visualize the data to understand the relationship of data and know the problem of data effectively. Then, we implement the data preprocessing to our data set which to make it become meaningful to Artificial Neural Network and Linear GAM. After that, we build the both of the model, and train the them by using the data had pre-processed. The prediction can be generated when the both of the model finished the training and make evaluation on both model to know the performance of model with the default value. We also will perform fine tune on both of the model, to let them can work in best statue and select the model as the main prediction model. At the last, the best performance between both of the model is the Linear GAM, so it will become the best model.

References

“Dengue and severe dengue,” *Who.int*. Accessed: March. 30, 2022.[Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>.

S. Prabhakaran, “ARIMA model - complete guide to time series forecasting in python,” *Machine Learning Plus*, 22-Aug-2021. Accessed: March. 30, 2022. [Online]. Available: <https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python>.

Dowshen.s“*Dengue Fever (for Teens) - Nemours KidsHealth*” *Kidshealth.org*. Accessed: March. 30, 2022.[Online]. Available: <https://kidshealth.org/en/teens/dengue.html>.

Norsyahida, “History and epidemiology of dengue,” *DENGGI*. Accessed: March. 31, 2022.[Online]. Available: <http://denggi.myhealth.gov.my/history-and-epidemiology-of-dengue/?lang=en>.

R. Jain, S. Sontisirikit, S. Iamsirithaworn, and H. Prendinger, “Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data,” *BMC Infect. Dis.*, vol. 19, no. 1, p. 272, 2019. Accessed: March. 31, 2022.[Online]. Available: <https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-019-3874-x>

A. E. Laureano-Rosario *et al.*, “Application of artificial neural networks for dengue fever outbreak predictions in the northwest coast of Yucatan, Mexico and San Juan, Puerto Rico,” *Trop. Med. Infect. Dis.*, vol. 3, no. 1, p. 5, 2018. Accessed: March. 31, 2022.[Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6136605/>

D. Liu *et al.*, “A dengue fever predicting model based on Baidu search index data and climate data in South China,” *PLoS One*, vol. 14, no. 12, p. e0226841, 2019. Accessed:

March. 31, 2022.[Online]. Available:
<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0226841>

Uniqtech, “Multilayer perceptron (MLP) vs convolutional neural network in Deep Learning,” *Data Science Bootcamp*, 22-Dec-2018. Accessed: April.1 , 2022.[Online]. Available: <https://medium.com/data-science-bootcamp/multilayer-perceptron-mlp-vs-convolutional-neural-network-in-deep-learning-c890f487a8f1>.

I. P.Pratama,“RELATIONSHIP BETWEEN URBANIZATION AND DENGUE HEMORRHAGIC FEVER INCIDENCE IN SEMARANG CITY”.Accessed: April. 1, 2022. [Online]. Available:[https://www.researchgate.net/publication/304069865_RELATIONSHIP_BE TWEEN_URBANIZATION_AND_DENGUE_HEMORRHAGIC_FEVER_INCIDEN CE_IN_SEMARANG_CITY](https://www.researchgate.net/publication/304069865_RELATIONSHIP_BE_TWEEN_URBANIZATION_AND_DENGUE_HEMORRHAGIC_FEVER_INCIDEN CE_IN_SEMARANG_CITY).

E. S. Shepard, E. A. Undurraga, R. S. Lees, Y. Halasa, L. C. S. Lum, and C. W. Ng, “Use of multiple data sources to estimate the economic cost of dengue illness in Malaysia,” *Am. J. Trop. Med. Hyg.*, vol. 87, no. 5, pp. 796–805, 2012.Accessed: April. 2, 2022.[Online].Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3516253/>

J. Hoare, “What is a random forest?,” *Displayr*, 07-May-2018.Accessed: April. 2, 2022. [Online]. Available: <https://www.displayr.com/what-is-a-random-forest/>.

S. K. Yamana, S. Kandula, and J. Shaman, “Superensemble forecasts of dengue outbreaks,” *J. R. Soc. Interface*, vol. 13, no. 123, 2016.Accessed: April. 2, 2022.[Online]. Available: [https://royalsocietypublishing.org/doi/10.1098/rsif.2016.0410#:~:text=Superensemble %20forecasts%20of%20peak%20timing,\(mean%20difference%206.9%20weeks\)](https://royalsocietypublishing.org/doi/10.1098/rsif.2016.0410#:~:text=Superensemble%20forecasts%20of%20peak%20timing,(mean%20difference%206.9%20weeks)).

N. Zhao *et al.*, “Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia,” *PLoS Negl. Trop. Dis.*, vol. 14, no. 9, p. e0008056, 2020.

Datacenterdynamics.com. Accessed: April. 2, 2022.[Online]. Available: <https://www.datacenterdynamics.com/en/opinions/three-reasons-to-store-historical-data/>.

O. -R. Adams, “How artificial intelligence works - becoming human: Artificial intelligence magazine,” *Becoming Human: Artificial Intelligence Magazine*, 31-Oct-2019. Accessed: April. 2, 2022.[Online]. Available: <https://becominghuman.ai/how-artificial-intelligence-currently-works-974e6782ddda>.

B. Kizil, “Three reasons to store historical data”. Accessed: April. 3, 2022.[Online]. Available: <https://www.datacenterdynamics.com/en/opinions/three-reasons-to-store-historical-data/>.

T. Mahidin, “Department of statistics Malaysia official portal,” *Gov.my*. Accessed: April.3,2022.[Online].Available:https://www.dosm.gov.my/v1/index.php?r=column/cthemByCat&cat=155&bul_id=aWJZRkJ4%20UEdKcUZpT2tVT090Snpdz09&menu_id=L0pheU43NWJwRWVSZklWdzQ4TlhUUT09.

Business Matters, “Jupyter Notebook for Data Science? The pros and cons of this tool,” *Business Matters*, 26-Feb-2021.Accessed: April. 3, 2022. [Online]. Available: <https://bmmagazine.co.uk/business/jupyter-notebook-for-data-science-the-pros-and-cons-of-this-tool/>.

“What is NumPy? — NumPy v1.22 Manual,” *Numpy.org*.Accessed: April. 4, 2022. [Online]. Available: <https://numpy.org/doc/stable/user/whatisnumpy.html>.

“What is Pandas in Python? Everything you need to know,” *ActiveState*, 09-Oct-2020. Accessed: April. 4, 2022.[Online]. Available: <https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/>.

“Matplotlib — Visualization with Python,” *Matplotlib.org*. Accessed: April. 4, 2022.[Online]. Available: <https://matplotlib.org/>.

K. Fernando, “R-Squared,” *Investopedia*, 08-Feb-2022. Accessed: April. 4, 2022. [Online]. Available: <https://www.investopedia.com/terms/r/r-squared.asp>.

Sciencedirect.com. Accessed: April. 4, 2022. [Online]. Available: <https://www.sciencedirect.com/topics/engineering/root-mean-squared-error>.

A. S. Masrani, N. R. Nik Husain, K. I. Musa, and A. S. Yasin, “Prediction of dengue incidence in the Northeast Malaysia based on weather data using the generalized additive model,” *Biomed Res. Int.*, vol. 2021, p. 3540964, 2021. Accessed: April. 5, 2022. [Online]. Available: <https://www.hindawi.com/journals/bmri/2021/3540964/>

Biomedcentral.com. Accessed: April. 5, 2022. [Online]. Available: <https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-019-3874-x#Sec7>.

N. A. M. Salim *et al.*, “Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques,” *Sci. Rep.*, vol. 11, no. 1, p. 939, 2021. Accessed: April. 5, 2022. [Online]. Available: <https://www.nature.com/articles/s41598-020-79193-2>

B. M. Sundram, D. B. Raja, F. Mydin, T. C. Yee, K. Raj, and F. Kamaludin, “Utilizing artificial intelligence as a dengue surveillance and prediction tool,” *Journal of Applied Bioinformatics & Computational Biology*, vol. 2019, 2019. Accessed: April. 5, 2022. [Online] Available: https://www.scitechnol.com/peer-review/utilizing-artificial-intelligence-as-a-dengue-surveillance-and-prediction-tool-P0vC.php?article_id=9445#:~:text=The%20system%20was%20developed%20primarily,single%20case%2C%20cluster%20or%20outbreaks.

Y. T. Choo *et al.*, “Artificial intelligence model as predictor for dengue outbreaks,” *Malays. J. Publ. Health Med.*, vol. 19, no. 2, pp. 103–108, 2019. Accessed: April. 5, 2022. [Online]. Available: <https://mjphm.org/index.php/mjphm/article/view/176#:~:text=Abstract,from%20the%20Ministry%20of%20Health.>

R. Jain, S. Sontisirikit, S. Iamsirithaworn, and H. Prendinger, “Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic

data,” *BMC Infect. Dis.*, vol. 19, no. 1, p. 272, 2019. Accessed: April. 6, 2022. [Online]. Available: <https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-019-3874-x>

Zulkifli, Zuliana, “Dengue Statistics in Malaysia, per state in 1999 to 2019”, Mendeley Data, V1, doi: 10.17632/r3k8rxg2cw.1

“Malaysia,” distributed by *Worldbank.org*. [Online]. Available: <https://climateknowledgeportal.worldbank.org/country/malaysia/climate-data-historical>

“Daily Projected Rainfall (CCSM3A1B) by State in Peninsular Malaysia - daily projected rainfall ccsm3a1b for 2020 by state in peninsular malaysia - MAMPU,” *Gov.my*. [Online]. Available: https://www.data.gov.my/data/ms_MY/dataset/daily-projected-rainfall-ccsm3a1b-by-state-in-peninsular-malaysia/resource/c324c4f1-7d41-4e68-8e94-e0f5ae5f6274

“dataset,” *Gov.my*. [Online]. Available: https://www.data.gov.my/data/ms_MY/dataset/population-by-state-administrative-district-and-sex/resource/184aee01-8562-42bc-9b05-75e601e445f0. [Accessed: 15-Apr-2022].

“States in Malaysia,” *Latlong.net*. [Online]. Available: <https://www.latlong.net/category/states-133-14.html>. [Accessed: 15-Apr-2022].

Medium. 2022. Introduction to Linear Regression and Polynomial Regression. [online] Available at: <https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb> [Accessed 9 September 2022].

C3 AI. 2022. Root Mean Square Error (RMSE) - C3 AI. [online] Available at: <https://c3.ai/glossary/data-science/root-mean-square-error-rmse/#:~:text=It%20shows%20how%20far%20predictions,square%20root%20of%20>

that%20mean.> [Accessed 9 September 2022].

Statistics How To. 2022. Absolute Error & Mean Absolute Error (MAE). [online] Available at: <<https://www.statisticshowto.com/absolute-error/>> [Accessed 9 September 2022].

Fonti Kar, a., 2022. Generalised Additive Models (GAMs) :: Environmental Computing.[online]Environmentalcomputing.net. Available at: <<https://environmentalcomputing.net/statistics/gams/#:~:text=Let's%20start%20with%20an%20equation,the%20linear%20predictor%20is%20now>> [Accessed 9 September 2022].

Analytics Vidhya. 2022. Estimation of Neurons and Forward Propagation in Neural Net. [online] Available at: <<https://www.analyticsvidhya.com/blog/2021/04/estimation-of-neurons-and-forward-propagation-in-neural-net/#:~:text=There%20are%20three%20steps%20to,loss%20or%20the%20error%20term.>> [Accessed 9 September 2022].

Weekly Log

**FINAL YEAR PROJECT WEEKLY REPORT
(Project II)**

Trimester, Year: Y4S1	Study week no.: 2
Student Name & ID: Cheo Jia Jun 1805681	
Supervisor: Dr. Chang Jing Jing	
Project Title: Development of Dengue Prediction Model With Neural Network	

1. WORK DONE 1. Meet With Supervisor 2. Schedule Project Plan 3. Collect Certain Research Material
2. WORK TO BE DONE 1. Analysis Research Material
2. PROBLEMS ENCOUNTERED None
4. SELF EVALUATION OF THE PROGRESS Smooth execute the Schedule



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT
(Project II)

Trimester, Year: Y4S1	Study week no.: 4
Student Name & ID: Cheo Jia Jun 1805681	
Supervisor: Dr. Chang Jing Jing	
Project Title: Development of Dengue Prediction Model With Neural Network	

1. WORK DONE

1. Analysis Research Material is still available or not

2. WORK TO BE DONE
Find the way to create GAM

2. PROBLEMS ENCOUNTERED
Lack of Data Source and Expensive of Data

4. SELF EVALUATION OF THE PROGRESS
Smoothly execute schedule



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT
(Project II)

Trimester, Year: Y4S1	Study week no.: 6
Student Name & ID: Cheo Jia Jun 1805681	
Supervisor: Dr. Chang Jing Jing	
Project Title: Development of Dengue Prediction Model With Neural Network	

<p>1. WORK DONE</p> <p>1. Create GAM</p>
<p>2. WORK TO BE DONE</p> <p>1. Train Artificial Neural Network and GAM</p> <p>2. Fine tune Model</p>
<p>2. PROBLEMS ENCOUNTERED</p> <p>None</p>
<p>4. SELF EVALUATION OF THE PROGRESS</p> <p>Smooth execute schedule</p>



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT
(Project I)

Trimester, Year: Y4S1	Study week no.: 8
Student Name & ID: Cheo Jia Jun 1805681	
Supervisor: Dr. Chang Jing Jing	
Project Title: Development of Dengue Prediction Model With Neural Network	

1. WORK DONE

- 1. Train Artificial Neural Network and GAM**
- 2. Fine tune Model**

2. WORK TO BE DONE

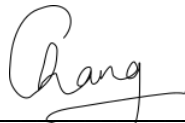
- 1. Conduct Evaluation of Model and make decision on the best model**
- 2. Prepare report**

2. PROBLEMS ENCOUNTERED

None

4. SELF EVALUATION OF THE PROGRESS

Smoothly execute schedule



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT
(Project II)

Trimester, Year: Y4S1	Study week no.: 10
Student Name & ID: Cheo Jia Jun 1805681	
Supervisor: Dr. Chang Jing Jing	
Project Title: Development of Dengue Prediction Model With Neural Network	

1. WORK DONE Prepare Report
3. WORK TO BE DONE Tunity Checking
2. PROBLEMS ENCOUNTERED None
4. SELF EVALUATION OF THE PROGRESS Smoothly execute schedule

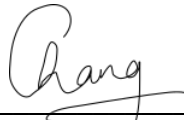
Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT
(Project II)

Trimester, Year: Y4S1	Study week no.: 12
Student Name & ID: Cheo Jia Jun 1805681	
Supervisor: Dr. Chang Jing Jing	
Project Title: Development of Dengue Prediction Model With Neural Network	

1. WORK DONE Tunity Checking
4. WORK TO BE DONE Project Submission
2. PROBLEMS ENCOUNTERED None
4. SELF EVALUATION OF THE PROGRESS Smoothly execute schedule √



Supervisor's signature



Student's signature



Faculty of Information and Communication Technology
 Name: Cheo Jia Jun
 Supervisor: Dr. Chang Jing Jing

Development of Dengue Prediction Model with Neural Network

Introduction

Since the 21st century, fighting various diseases or viruses has become an important part of human beings. The most common disease among them is dengue fever. Dengue is a mosquito-borne viral infection. dengue not only will affect humans health, at the same time it will affect the economy. an adjusted estimate of economic burden due to dengue illness is RM196 million per year, which is approximately RM7.14 per capital. In this project, we will develop a system with high accuracy to predict Dengue outbreak which in turn help to decrease the dengue case by taking necessary precaution. The results from Artificial Neural Network (ANN) and Generalized Additive Models (GAMs) will be compared.

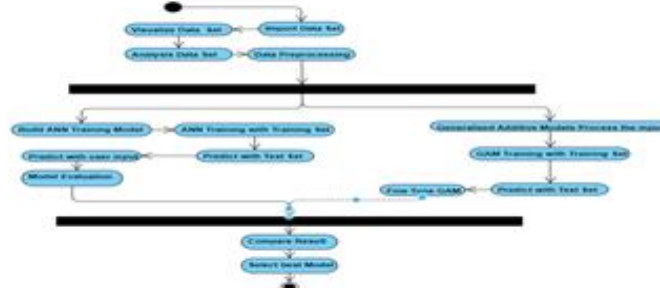
Objectives

- ✓ To identify prediction variables that will impact dengue outbreak.
- ✓ To develop neural network models to predict the number of dengue incidence in Malaysia.
- ✓ To compare the performance of ANN and GAM

Project Scopes

- ✓ Prediction model of dengue fever infections among Malaysians
- ✓ Comparing the accuracy of prediction result of different prediction models
- ✓ To effectively control the mosquito breeding grounds for effective actions

Methodology



Results

Artificial Neural Network
 Mean Absolute Error: 2523.31
 Root Mean Square Error: 3372.39
 R2: 92.54%

Generalize Additive Model
 Mean Absolute Error: 1704.48
 Root Mean Square Error: 2043.43
 R2: 97.26%

Test True Y	Model Predictions
0	48349
1	423
2	8071
3	12873
4	239
5	12641
6	2140
7	384
8	852
9	7932
10	3884

After perform the evaluation and comparison on ANN and GAM, the GAM is the model that most suitable to predict the outbreak of dengue as Root Mean Squared Error and Mean Absolute Error is less than ANN which mean the prediction is close to the best fit line. Besides that, the R2 of the GAM also is higher than ANN which the prediction is fit to the model. The prediction of GAM also very close to the actual data.

PLAGIARISM CHECK RESULT

12%

SIMILARITY INDEX

%

INTERNET SOURCES

12%

PUBLICATIONS

%

STUDENT PAPERS

PRIMARY SOURCES

- 1** Naizhuo Zhao, Katia Charland, Mabel Carabali, Elaine O. Nsoesie et al. "Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia", PLOS Neglected Tropical Diseases, 2020 1%
Publication

- 2** Raghvendra Jain, Sra Sontisirikit, Sapon Iamsirithaworn, Helmut Prendinger. "Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data", BMC Infectious Diseases, 2019 1%
Publication

- 3** Nurul Azam Mohd Salim, Yap Bee Wah, Caitlynn Reeves, Madison Smith et al. "Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques", Scientific Reports, 2021 1%
Publication

4	Patsaraporn Somboonsak. "Development Innovation to Predict Dengue Affected Area and Alert People with Smartphones", International Journal of Online and Biomedical Engineering (IJOE), 2020 <small>Publication</small>	1%
5	Teresa K. Yamana, Sasikiran Kandula, Jeffrey Shaman. "Superensemble forecasts of dengue outbreaks", Journal of The Royal Society Interface, 2016 <small>Publication</small>	1%
6	Afiqah Syamimi Masrani, Nik Rosmawati Nik Husain, Kamarul Imran Musa, Ahmad Syaarani Yasin. "Prediction of Dengue Incidence in the Northeast Malaysia Based on Weather Data Using the Generalized Additive Model", BioMed Research International, 2021 <small>Publication</small>	1%
7	Dan Liu, Songjing Guo, Mingjun Zou, Cong Chen, Fei Deng, Zhong Xie, Sheng Hu, Liang Wu. "A dengue fever predicting model based on Baidu search index data and climate data in South China", PLOS ONE, 2019 <small>Publication</small>	1%
8	P. Siriyasatien, S. Chadsuthi, K. Jampachaisri, K. Kesorn. "Dengue Epidemics Prediction: A Survey of the State-of-the-Art based on Data Science Processes", IEEE Access, 2018 <small>Publication</small>	<1%

9	Abdiel Laureano-Rosario, Andrew Duncan, Pablo Mendez-Lazaro, Julian Garcia-Rejon et al. "Application of Artificial Neural Networks for Dengue Fever Outbreak Predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico", Tropical Medicine and Infectious Disease, 2018 <small>Publication</small>	<1%
10	Xin-She Yang. "Mathematical foundations", Elsevier BV, 2019 <small>Publication</small>	<1%
11	"Pattern Recognition Applications and Methods", Springer Science and Business Media LLC, 2015 <small>Publication</small>	<1%
12	Cong-Han Zheng, Ping-Yu Hsu, Ming-Shien Cheng, Ni Xu, Yu-Chun Chen. "Chapter 7 Predicting Infection Area of Dengue Fever for Next Week Through Multiple Factors", Springer Science and Business Media LLC, 2022 <small>Publication</small>	<1%
13	"Intelligent Systems", Springer Science and Business Media LLC, 2022 <small>Publication</small>	<1%
14	Lecture Notes in Computer Science, 2007. <small>Publication</small>	<1%

- 15** Ashraf Kadry, Sumner Norman, Jing Xu, Deborah Solomonow-Avnon, Firas Mawase. "Computational neural network provides naturalistic solution for recovery of finger dexterity after stroke", Cold Spring Harbor Laboratory, 2021
Publication <1%
-
- 16** Merve Erkinay Özdemir, Ziya Telatar, Osman Eroğul, Yusuf Tunca. "Classifying dysmorphic syndromes by using artificial neural network based hierarchical decision tree", Australasian Physical & Engineering Sciences in Medicine, 2018
Publication <1%
-
- 17** Wiwik Anggraeni, Surya Sumpeno, Eko Mulyanto Yuniarno, Reza Fuad Rachmadi, Agustinus Bimo Gumelar, Mauridhi H Purnomo. "Prediction of Dengue Fever Outbreak Based on Climate Factors Using Fuzzy-Logistic Regression", 2020 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2020
Publication <1%
-
- 18** *Studies in Computational Intelligence*, 2010.
Publication <1%
-
- 19** Lu-Ping Luo, Chao Yuan, Rui-Jun Yan, Quan Yuan, Jing Wu, Kyoo-Sik Shin, Chang-Soo Han. "Trajectory planning for energy minimization <1%

of industry robotic manipulators using the Lagrange interpolation method", International Journal of Precision Engineering and Manufacturing, 2015

Publication

20 "Assessing COVID-19 and Other Pandemics and Epidemics using Computational Modelling and Data Analysis", Springer Science and Business Media LLC, 2022

Publication

21 Elmer Andres Fernandez. "Comparison of Standard and Artificial Neural Network Estimators of Hemodialysis Adequacy", Artificial Organs, 2/2005

Publication

22 Xin Li, Lingzhe Zhang, Chenyu Lin, Jingyao Chen, Jun Li, Yuqian Zhu, Jinfeng Zhu, Lijin Chen, Xiaofeng Li, Bingdi Chen. "Learning an Improved Object Detection Approach Based on the YOLO Algorithm to Identify Circulating Tumor Cells", Research Square Platform LLC, 2022

Publication

23 Lathesparan Ramachandran, Rm Kapila Tharanga Rathnayaka, Wiraj Udara Wickramaarachchi. "Artificial Neural Networks Based Dengue Diagnosis Prediction Model", 2021 IEEE 16th International Conference on

Industrial and Information Systems (ICIIS),
2021

Publication

24 "Fourth International Congress on
Information and Communication Technology",
Springer Science and Business Media LLC,
2020

Publication

25 S Ruban, Sanjeev Rai. "Enabling data to
develop an AI-based application for detecting
malaria and dengue", Walter de Gruyter
GmbH, 2021

Publication

26 Madeleine Charney. "
<http://sdwebx.worldbank.org/climateportal> ",
Journal of Agricultural & Food Information,
2017

Publication

27 Mahmood Akhtar, Moritz U. G. Kraemer,
Lauren M. Gardner. "A dynamic neural
network model for predicting risk of Zika in
real time", BMC Medicine, 2019

Publication

28 Naizhuo Zhao, Katia Charland, Mabel Carabali,
Elaine Nsoesie et al. "Machine learning and
dengue forecasting: Comparing random
forests and artificial neural networks for
predicting dengue burdens at the national

33 Nittaya Kerdprasop, Kittisak Kerdprasop, Paradee Chuaybamroong. "Computational Intelligence and Statistical Learning Performances on Predicting Dengue Incidence using Remote Sensing Data", Advances in Science, Technology and Engineering Systems Journal, 2020

<1%

Publication

34 Rafael Bomfim, Sen Pei, Jeffrey Shaman, Teresa Yamana, Hernán A. Makse, José S. Andrade, Antonio S. Lima Neto, Vasco Furtado. "Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas", Journal of The Royal Society Interface, 2020

<1%

Publication

35 Xia-ting Feng, M. Seto, K. Katsuyama. "Neural dynamic modelling on earthquake magnitude series", Geophysical Journal International, 1997

<1%

Publication

FYP2 CHECKLIST

Universiti Tunku Abdul Rahman			
Form Title : Supervisor’s Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective	Date
			Page No.: 1of 1



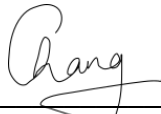
**FACULTY OF INFORMATION AND COMMUNICATION
TECHNOLOGY**

Full Name(s) of Candidate(s)	Cheo Jia Jun
ID Number(s)	18ACB05681
Programme / Course	IA
Title of Final Year Project	Development of Dengue Prediction Model With Neural Network

Similarity	Supervisor’s Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)
Overall similarity index: <u>12</u> % Similarity by source Internet Sources:0% Publications:12% Student Papers:0%	The low similiary index indicates a high originality of this report.
Number of individual sources listed or more than 3% similarity: <u>0%</u>	-
Parameters of originality required and limits approved by UTAR are as Follows: (i) Overall similarity index is 20% and below, and (ii) Matching of individual sources listed must be less than 3% each, and (iii) Matching texts in continuous block must not exceed 8 words <i>Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.</i>	

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.



Signature of Supervisor

Name: Chang Jing Jing

Date:

12/9/2022

Signature of Co-Supervisor

Name:

Date:



UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY
(KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	18ACB05681
Student Name	Cheo Jia Jun
Supervisor Name	Ts Dr.Chang Jing Jing

TICK (√)	DOCUMENT ITEMS
	Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
√	Front Plastic Cover (for hardcopy)
√	Title Page
√	Signed Report Status Declaration Form
√	Signed FYP Thesis Submission Form
√	Signed form of the Declaration of Originality
√	Acknowledgement
√	Abstract
√	Table of Contents
√	List of Figures (if applicable)
√	List of Tables (if applicable)
√	List of Symbols (if applicable)
√	List of Abbreviations (if applicable)
√	Chapters / Content
√	Bibliography (or References)
√	All references in bibliography are cited in the thesis, especially in the chapter of literature review
√	Appendices (if applicable)
√	Weekly Log
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)
√	I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report.

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.

(Signature of Student)

Date:

