

Predictive Analysis of Malaysian Dengue Hemorrhagic Fever data from 2010 - 2017 Using BigML

Foong M. Wong

November 15, 2019

- Dengue Hemorrhagic Fever (DHF)
- Dengue Symptoms
- Dengue Dataset from Malaysian Open Data Portal

The screenshot displays the Malaysian Open Data Portal interface. The breadcrumb trail at the top reads: Home / Data Set Provider / Ministry of Health / Number of Dengue Disease Cases ...

On the left sidebar, there is a search bar with the text 'Carian Terperinci' and a 'LOOK FOR' button. Below this, the 'Data Set Provider' section features the Malaysian coat of arms and the text 'health Ministry'. Further down, there are sections for 'Openness' (indicated by an eye icon), a star rating (★★☆☆☆), 'Social' media links, and 'Google+'.

The main content area has a search bar with the text 'Carl' and a 'LOOK FOR' button. Below the search bar, there are tabs for 'Data Set' (selected) and 'Clusters'. The title of the dataset is 'Number of Annual Dengue Haemorrhagic Fever (DHF) Cases by State and Age'. The description states: 'This dataset shows the number of cases of Dengue Haemorrhagic Fever (DHF) cases annually by state and age in Malaysia. Data will be provided annually once the data is finalized.' It also shows 'Total Views: 0' and 'Status: Dataset is Published'.

Under the 'Data Sets and Sources' section, there is a preview of the dataset with a green 'X' icon, the title 'Number of cases of Dengue Haemorrhagic Fever Disease ...', a description 'This dataset shows the number of cases of Dengue Haemorrhagic Fever (DHF) ...', and 'Downloads: 6'. An 'Explore' button is located to the right of this section.

Number of Annual DHF Cases by State and Age from 2010 - 2017

Recap - Research Goals

- Analyze the historical (based on 2010-2017 data from the portal) trend of:
 - ① Age groups of dengue hemorrhagic fever (today)
 - ② Infected areas/states (today)
- Run predictive/classification analysis describe:
 - ① Epidemic pattern (future)
 - ② Future outbreaks (future)

- ① Data Formatting (Python)
- ② Data Loading (BigML)
- ③ Data Merging (BigML)
- ④ Data Overview (BigML)

Data Formatting

Original Data

BILANGAN KES PENYAKIT DENGUE HAEMORRHAGIC FEVER TAHUNAN MENGIKUT NEGERI DAN KUMPULAN UMUR, TAHUN 2010																
KUMPULAN UMUR	JOHOR	KEDAH	KELANTAN	MELAKA	NEGERI SEMBILAN	PAHANG	PERAK	PERLIS	PULAU PINANG	SABAH	SARAWAK	SELANGOR	TERENGGANU	WP KUALA LUMPUR	WP LABUAN	MALAYSIA
0-4	10	0	4	1	0	0	0	0	3	1	5	32	2	5	0	63
5-9	16	0	11	2	2	4	4	0	1	3	12	63	4	21	0	143
10-14	63	0	25	29	9	12	8	0	5	2	10	182	9	37	0	391
15-19	67	0	55	38	15	16	4	0	1	2	12	243	14	38	0	505
20-24	102	0	32	52	18	10	2	0	0	0	22	348	13	44	0	643
25-29	115	0	31	41	18	13	2	0	3	2	21	330	14	48	0	638
30-34	78	0	20	28	14	13	4	0	2	2	12	246	5	33	0	457
35-39	63	0	23	28	13	10	1	0	4	1	17	166	6	20	0	352
40-44	35	1	14	27	9	7	2	0	1	3	20	132	5	15	0	271
45-49	29	1	22	18	13	12	0	0	1	1	8	83	3	6	0	197
50-54	18	0	21	15	9	11	1	0	0	0	11	61	3	11	0	161
55-59	12	1	6	16	6	4	3	0	0	0	4	34	4	10	0	100
60-64	12	0	5	5	2	2	0	0	0	0	3	16	0	5	0	50
65-69	7	0	8	3	0	1	0	0	0	0	3	10	0	2	0	34
70-74	3	0	0	4	0	0	1	0	0	0	0	7	0	2	0	17
>75	0	0	2	1	0	1	0	0	0	0	2	2	1	0	0	9
Grand Total	630	3	279	308	128	116	32	0	21	17	162	1955	83	297	0	4031

Dengue Hemorrhagic Fever 2010 Dataset

Data Formatting (cont.)

- Include cases from 13 states, remove cases from 2 federal territories
- Avoid duplicated cases (since 2 federal territories are part of the 13 states)

BILANGAN KES PENYAKIT DENGUE HAEMORRHAGIC FEVER TAHUNAN MENGIKUT NEGERI DAN KUMPULAN UMUR, TAHUN 2010																
KUMPULAN UMUR	JOHOR	KEDAH	KELANTAN	MELAKA	NEGERI SEMBILAN	PAHANG	PERAK	PERLIS	PULAU PINANG	SABAH	SARAWAK	SELANGOR	TERENGGANU	WP KUALA LUMPUR	WP LABUAN	MALAYSIA
0-4	10	0	4	1	0	0	0	0	3	1	5	32	2	5	0	63
5-9	16	0	11	2	2	4	4	0	1	3	12	63	4	21	0	143
10-14	63	0	25	29	9	12	8	0	5	2	10	182	9	17	0	391
15-19	67	0	55	38	15	16	4	0	1	2	12	243	14	8	0	505
20-24	102	0	32	52	18	10	2	0	0	0	22	348	13	4	0	643
25-29	115	0	31	41	18	13	2	0	3	2	21	330	14	4	0	638
30-34	78	0	20	28	14	13	4	0	2	2	12	246	5	3	0	457
35-39	63	0	23	28	13	10	1	0	4	1	17	166	6	2	0	352
40-44	35	1	14	27	9	7	2	0	1	3	20	132	5	5	0	271
45-49	29	1	22	18	13	12	0	0	1	1	8	83	3	6	0	197
50-54	18	0	21	15	9	11	1	0	0	0	11	61	3	11	0	161
55-59	12	1	6	16	6	4	3	0	0	0	4	34	4	10	0	100
60-64	12	0	5	5	2	2	0	0	0	0	3	16	0	5	0	50
65-69	7	0	8	3	0	1	0	0	0	0	3	10	0	2	0	34
70-74	3	0	0	4	0	0	1	0	0	0	0	7	0	2	0	17
>75	0	0	2	1	0	1	0	0	0	0	2	2	1	0	0	9
Grand Total	630	3	279	308	128	116	32	0	21	17	162	1955	83	297	0	4031

Dengue Hemorrhagic Fever 2010 Dataset

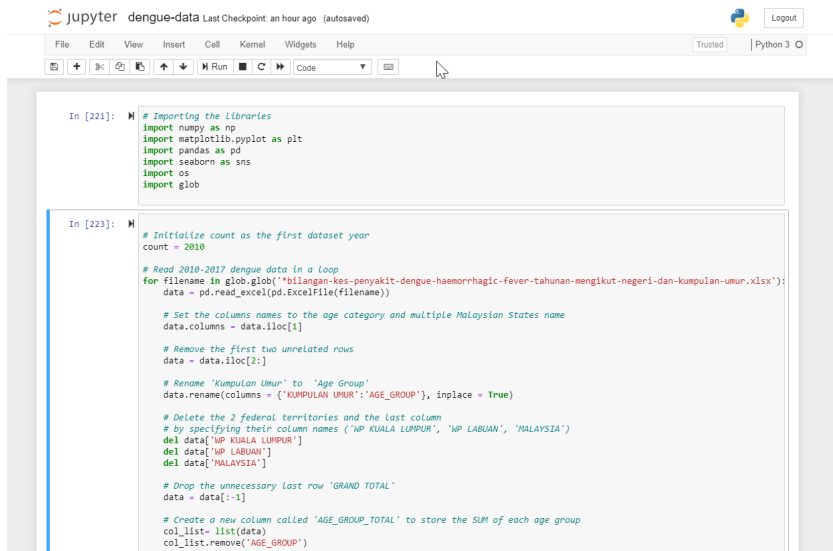
Data Formatting (cont.)

- Remove the last column and last row
- Rename 'KUMPULAN UMUR' to 'AGE_GROUP'

KUMPULAN UMUR	JOHOR	KEDAH	KELANTAN	MELAKA	NEGERI SEMBILAN	PAHANG	PERAK	PERLIS	PULAU PINANG	SABAH	SARAWAK	SELANGOR	TERENGGANU	WP KUALA LUMPUR	WP LAEUN	MALAYSIA
0-4	10	0	4	1	0	0	0	0	3	1	5	32	2	5	0	63
5-9	16	0	11	2	2	4	4	0	1	3	12	63	4	21	0	143
10-14	63	0	25	29	9	12	8	0	5	2	10	182	9	27	0	391
15-19	67	0	55	38	15	16	4	0	1	2	12	243	14	8	0	505
20-24	102	0	32	52	18	10	2	0	0	0	22	348	13	4	0	643
25-29	115	0	31	41	18	13	2	0	3	2	21	330	14	4	0	638
30-34	78	0	20	28	14	13	4	0	2	2	12	246	5	3	0	457
35-39	63	0	23	28	13	10	1	0	4	1	17	166	6	2	0	352
40-44	35	1	14	27	9	7	2	0	1	3	20	132	5	3	0	171
45-49	29	1	22	18	13	12	0	0	1	1	8	83	3	6	0	197
50-54	18	0	21	15	9	11	1	0	0	0	11	61	3	11	0	161
55-59	12	1	6	16	6	4	3	0	0	0	4	34	4	10	0	100
60-64	12	0	5	5	2	2	0	0	0	0	3	16	0	5	0	60
65-69	7	0	8	3	0	1	0	0	0	0	3	10	0	2	0	31
70-74	3	0	0	4	0	0	1	0	0	0	0	7	0	2	0	11
>75	0	0	2	1	0	1	0	0	0	0	2	2	1	0	0	9

Dengue Hemorrhagic Fever 2010 Dataset

Data Formatting (cont.)



```
In [221]: # Importing the Libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import os
import glob

In [223]: # Initialize count as the first dataset year
count = 2010

# Read 2010-2017 dengue data in a loop
for filename in glob.glob('*bilangan-Kes-penyakit-dengue-haemorrhagic-fever-tahunan-mengikut-negeri-dan-kumpulan-umur.xlsx'):
    data = pd.read_excel(pd.ExcelFile(filename))

    # Set the columns names to the age category and multiple Malaysian States name
    data.columns = data.iloc[1]

    # Remove the first two unrelated rows
    data = data.iloc[2:]

    # Rename 'Kumpulan Umur' to 'Age Group'
    data.rename(columns = {'KUMPULAN UMUR': 'AGE_GROUP'}, inplace = True)

    # Delete the 2 federal territories and the last column
    # by specifying their column names ('WP KUALA LUMPUR', 'WP LABUAN', 'MALAYSIA')
    del data['WP KUALA LUMPUR']
    del data['WP LABUAN']
    del data['MALAYSIA']

    # Drop the unnecessary last row 'GRAND TOTAL'
    data = data[:-1]

    # Create a new column called 'AGE_GROUP_TOTAL' to store the SUM of each age group
    col_list = list(data)
    col_list.remove('AGE_GROUP')
```

Python script to create formatted datasets to be read into BigML

Data Loading

- Online machine learning service
- Easy to use

The screenshot displays the BigML web interface. At the top, the navigation bar includes the BigML logo, links for FEATURES, GALLERY, PRICING, and LABS (with a 'NEW' badge), and a 'Dashboard' button. A promotional banner for a 'FREE PRO Subscription' for students and educators is visible. The main heading reads 'Start making Data-driven Decisions today!' with the subtext 'No more wildly expensive or painful solutions'. A large green button labeled 'open dashboard' is centered below the heading. The bottom section shows a collage of four dashboard panels: 'Potential Fraudsters sample' with a scatter plot, 'Next Best Offer' with a bubble chart, 'Churn in Telephony' with a decision tree, and 'Anomaly assembly detector' with a data distribution plot. The text 'Machine Learning made easy, beautiful and understandable' is at the bottom of the collage.

BigML: A machine learning platform

Data Loading (cont.)

ml PRODUCT GETTING STARTED PRICING SUPPORT FOONGMINWONG Dashboard

FOONGMINWONG - My Dashboard Malaysian Dengue Hemorrhagic Fever 2010-2017

Sources Datasets Supervised Unsupervised Predictions Tasks WhizzML

Type	Name	Fields	Time	Size	Count
XL.S	dengue2014.xlsx	15 fields (1 categorical, 14 numeric)	1h 56min	737 bytes	1
XL.S	dengue2015.xlsx	15 fields (1 categorical, 14 numeric)	1h 56min	733 bytes	1
XL.S	dengue2017.xlsx	15 fields (1 categorical, 14 numeric)	2h 6min	724 bytes	1
XL.S	dengue2016.xlsx	15 fields (1 categorical, 14 numeric)	2h 6min	727 bytes	1
XL.S	dengue2013.xlsx	15 fields (1 categorical, 14 numeric)	2h 6min	733 bytes	1
XL.S	dengue2011.xlsx	15 fields (1 categorical, 14 numeric)	2h 6min	747 bytes	1
XL.S	dengue2012.xlsx	15 fields (1 categorical, 14 numeric)	2h 6min	725 bytes	1
XL.S	dengue2010.xlsx	15 fields (1 categorical, 14 numeric)	2h 9min	808 bytes	2

Show 10 sources 1 to 8 of 8 sources

Loading 2010-2017 datasets into BigML

Data Merging



bigml.com/dashboard/dataset/5dcdd0aff80b1640d7004a90/table



MERGE DATASETS CONFIGURATION

Current dataset:
dengue2010

Merge the current dataset with:

dengue2011

dengue2012

dengue2013

dengue2014

dengue2015

dengue2016

dengue2017

+ Add dataset

Dataset name:
dengue2010-2017

Sample rate: 100%

Sample rate: 100%

Sample rate: 100%

Sample rate: 100%

Sample rate: 100%

Sample rate: 100%

Sample rate: 100%

Sample rate: 100%

Seed

Seed

Seed

Seed

Seed

Seed

Seed

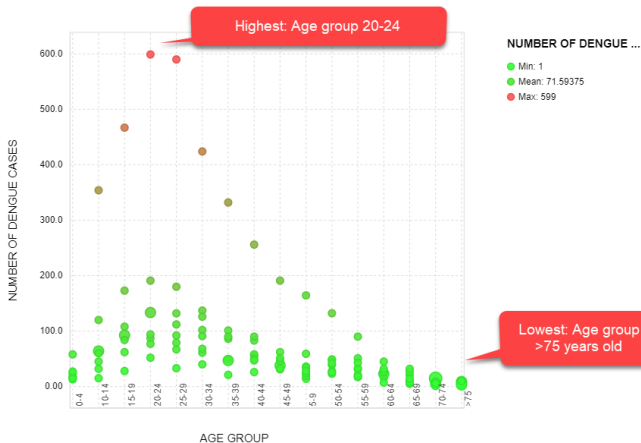
Seed

Reset

Merge datasets

Merging 2010 - 2017 individual datasets

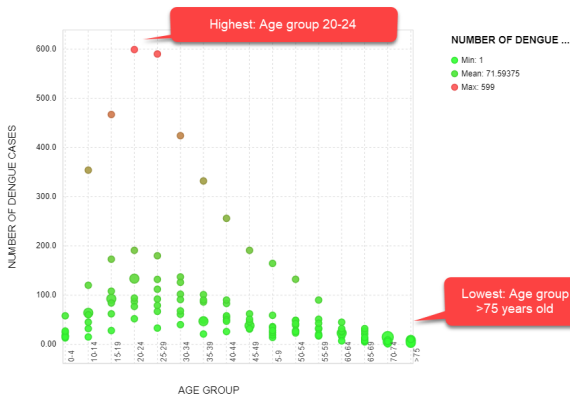
Data Overview - Age Groups



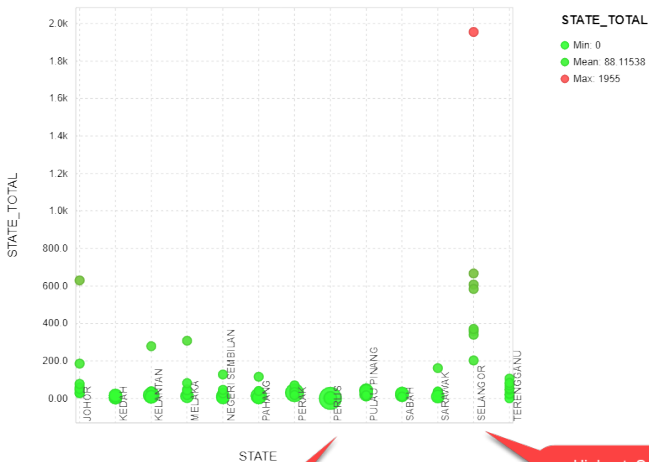
*Age Group with the Most number of DHF cases: (20-24),
Age Group with the Least number of DHF cases: (> 75)*

Data Overview - Age Groups (cont.)

- In South-East Asia, severe DHF cases are predominant among paediatric patients aged between 2 and 15 years (R. Bhatia, A.P. Dash, T. Sunyoto, 2013)
- In Malaysia, majority of the affected community are the age group of 13–35 years old (Sam et al, 2013)



Data Overview - 13 states



Lowest: Perlis

Highest: Selangor

State with Most number of DHF cases: Selangor, Least: Perlis

- ✓ Pre-process data to get into the right format
- ✓ Use BigML to plot 2010 - 2017 dengue hemorrhagic fever data
- ✗ Use BigML to conduct predictive analysis to forecast future outbreaks based 2010 - 2017 historical dengue data
- ✗ Choose what model to use: Decision tree? Time Series? Others?

References



Datasets of Number of Annual Dengue Hemorrhagic Fever (DHF) Cases by State and Age

<http://www.data.gov.my/data/MY/dataset/bilangan-kes\ -penyakit-dengue-haemorrhagic-fever-dhf-tahunan-mengikut\ -negeri-dan-umur. 2010 - 2017.>



BigML.com: A comprehensive Machine Learning platform for all predictive use cases. <https://bigml.com/>. 2019.



Data Formatting Python Script

<https://github.com/foongminwong/dengue-analysis/blob/master/dengue-data.ipynb>. 2015.



R. Bhatia, A.P. Dash, T. Sunyoto Changing epidemiology of dengue in South-East Asia WHO South-East Asia J Public Health, 2 (1) (2013), pp. 23-27



S.S. Sam, S.F. Omar, B.T. Teoh, J. Abd-Jamil, S. AbuBakar Review of dengue hemorrhagic fever fatal cases seen among adults: a retrospective study PLoS Negl Trop Dis, 7 (5) (2013), p. e2194