An Innovative Data Mining and Dashboard System for Monitoring of Malaysian Dengue Trends

Jastini Mohd Jamil, Izwan Nizal Mohd Shaharanee and Ve Chun Yung
School of Quantitative Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah.
jastini@uum.edu.my

***Abstract*—Monitoring dengue fever become an important task in reducing dengue outbreaks crisis. These monitoring tasks offered the stakeholder such as the Ministry of Health Malaysia (MOH) well informed status of the dengue fever. There are abundant dengue cases reported in Malaysia including mortality recorded over the past year. Data from Malaysian Open Data portal reveals, 21,900 cases of dengue fever were reported in 2012 with 35 deaths. However, this information are dispersed and circulated among several ministry and stakeholder. As such, information regarding the dengue outbreak belongs to MOH, while the information of population and density belong to another stakeholder. Putting this information into one monitoring system required an innovative system that capable to extract many data and information from several databases and capable to summarize these data into meaningful information. Knowing the dangerous effect of dengue fever, thus one of the solutions is to implement an innovative forecasting and dashboard system of dengue spread in Malaysia, with emphasize on an early prediction of dengue outbreak. Importantly, this research will deliver the message to health policy makers such as The Ministry of Health Malaysia (MOH), practitioners, and researchers of the importance to integrate their collaboration in exploring the potential strategies in order to reduce the future burden of the increase in dengue transmission cases in Malaysia.**

***Index Terms*—Dengue Fever Monitoring; Dashboard System; Knowledge Discovery; Data Mining.**

I. INTRODUCTION

A soaring number of dengue fever (DF) and dengue hemorrhagic fevers (DHF) have been recorded in Malaysia for the past several years [1]. These mosquito related fever continue to be an important public health problem in Malaysia. From 2010 to 2015, a large number of dengue fever cases were recorded by Ministry of Health, Vector Borne Disease Control Section (VBDC). In comparison with the neighborhood countries especially ASEAN, Malaysia was reported to have, higher case fatality rates (4.67%) compared with the neighboring countries like Thailand and Indonesia, with the case fatality rates of 0.3% and 0.5% respectively. While Malaysian has a decent surveillance system for dengue monitoring system, however, it is a passive system and has a little predictive capability [2]. Problem may occur if one waits for laboratory confirmation of the case before notification. Delay in notification may lead to delay in control measure, which will further lead to occurrence of outbreaks, since dengue needs optimum time of management as the transformation of DF into severe form of dengue are only takes a very short period. One of the solutions is to implement an innovative data mining and dashboard system of dengue spread in Malaysia, with emphasize on an early prediction of dengue outbreak using sophisticated data mining approach [2]. The data mining and dashboard system could improve public health problem with related to dengue fever in Malaysia since the accurate and well-summarized information to predict the dengue outbreak. This will enable timely action by public health officials to control such epidemics and mitigate their impact on human health [3]. Therefore, developing of dengue break prediction using data mining and dashboard system that incorporates location, time and intensity are needed to help and produce a prediction for early identified the specific location, temporal and intensity of dengue break accurately [4].

II. RELATED WORKS

Many researches have been conducted to analyze the dengue fever and its trends. [5] and [4] had implemented the Neural Network Model and Hidden Markov Model in dengue outbreak prediction. [6] had implemented the Autoregressive Integrated Moving Average (ARIMA) Model where the models were analyzes with the Box-Jenkins approach which was appropriate for a long forecasting period. This method for selecting an appropriate ARIMA model for estimating and forecasting a univariate time-series consisted of identification, estimation, diagnostic checking and forecasting. Mathematical models such as Susceptible-Infective-Susceptible (SIS) and Susceptible-Infective-Removed (SIR) has been utilized by [7] to predict the dengue fever incidence in Rio de Janeiro. Additionally, [8] implemented the forecasting models using climate variables as predictors for the time series analysis of dengue incidence in Guadeloupe.

From the above discussion on the DF and DHF and from the definition of dengue outbreak, we intend to employ this data mining and dashboard system for dengue disease outbreak monitoring system.

III. RESEARCH FRAMEWORK

This study uses data mining and dashboard application to build an innovative dengue fever (DF) and dengue hemorrhagic fevers (DHF) monitoring system in Malaysia. Data mining is the use of algorithms to extract the information and patterns from the database by the Knowledge Discovery Process (KDD) process. This process applies algorithms to transform the data and generate the desired results. This study consists of four main phases as depicted in Figure 1. The first phase is the data selection, which includes identifying the data related to dengue

fever from The Public Sector Open Data Portal (data.gov.my). The second phase is the preprocessing phase. In this phase, any necessary pre-processing is applied to the selected data, to ensure clean and consistent data. The third phase involves developing a profiles and data summarization towards the dengue data obtained from the previous phase. Here, a visualization of dengue fever status and trends will be summarizes using Microsoft Excel dashboard system.

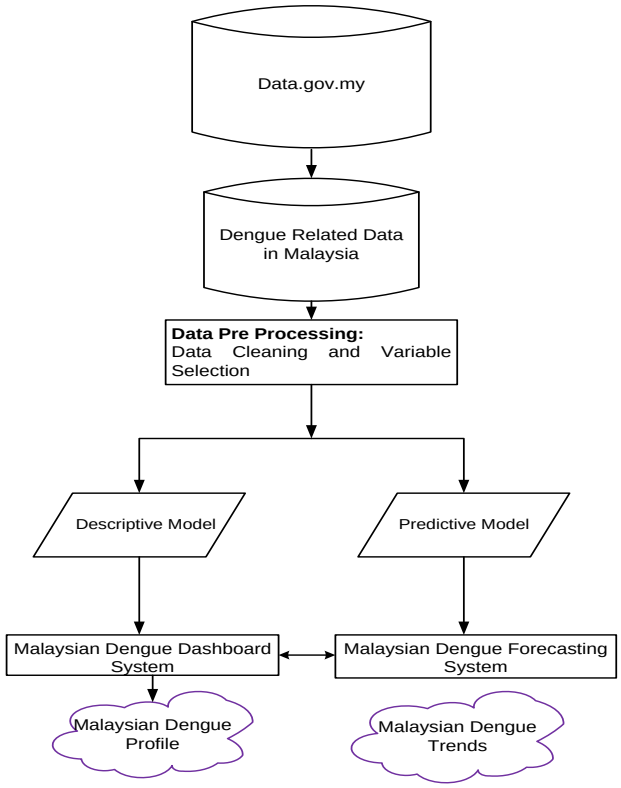


Figure 1: Data Mining and Dashboard System Framework for Monitoring of Malaysian Dengue Trends

A. *Phase 1: Data Selection*

Selecting an appropriate data set is highly important for the success of the project. The data source must obtain from the reliable and trustful source. Therefore, the data source for this project was obtained from The Public Sector Open Data Portal (data.gov.my). The portal serves as an online one-service-center to access and download the open government data. These data enables the government to share their data to a wide range of users and to increase the transparency of government services. The portal provides opportunities to citizens and the business community to increase their creativity and innovation in the creation of new products and offer a platform for the people to obtain information from official sources of the Government and as a means to obtain feedback from citizens. The data related to this project selected were from year 2010 to 2015 (September) by state. The variable selected is the number of dengue cases happened by state from year 2010 to 2015.

B. *Phase 2: Data Preprocessing*

After selecting the data, pre-processing step will be done. The main aim of pre-processing is to ensure the data is clean and possess high quality. Data pre-processing can be done using several methods such as data cleaning, data integration, data transformation, data reduction and data discretization. Data cleaning is a method to handle the incomplete, noisy and inconsistent data. Incomplete data can be done by estimating the probability value using regression. On the other hand, noisy data, which is caused by the variance in data, can be solved using outlier removal method or binning method. Data integration is done to compile the inconsistencies data, which comes from different sources and different naming standard. It can be done by consolidating different data into one repository. Besides that, correlation analysis can be done too to measure the strength of the relationship between different attribute. Data reduction is done to increase efficiency and reduce the huge data set into a smaller representation. Data discretization is used to transform the numerical value to categorical values. The target data set contains missing value, inconsistencies, redundant, and irrelevant values. It is essential to remove these inconsistencies, noise and outliers. Here, the pre-processing is applied to each attribute/s in the dataset in order to obtain clean and consistent data.

C. *Phase 3: Descriptive Model using Data Visualization Techniques*

Data summarizations are done on the attributes in the dataset. The data descriptions are done based on the number frequency of dengue cases happened by years and states. The main purpose of the data summarization is to allow the users to have clearer pictures for the dengue outbreak at Malaysia in order for them could have precise target to solve the dengue outbreak issues. Additionally, the dashboard system developed could be customized in terms of users and expectations. This allows each person to see the level of detail that they need in order to get their job done and meet their goals.

With refer to the data.gov.my portal; there are many data sources available for monitoring dengue data. Users would spend large amount of time reviewing and analyzing different datasets/reports to end in a conclusion. This dashboard tool allows to see, at a glance, an overall situation report of the desired information. This dashboards system is developed with the ability to get as deeper in information as required by simply selecting the desired variable or object. The graphical design allows an easy and smooth navigation throughout the information.

D. *Phase 4: Predictive Model using Data Mining Techniques*

The main aim of predictive modelling is to make predictions about values of data using known results or based on other historical data. Several techniques for predictive modelling tasks are classification, prediction and outlier detection. [7] agree that the function of the prediction model in the data mining task is to determine the future/new outcome(s) rather than discover current behaviours. Additionally, the outcomes of the prediction model may be either categorical or numerical values as opposed to those of the classification task.

Time-series approach such as Exponential Smoothing, Moving Average and Simple Linear Regression approaches have been implemented in this project in order to let user to do the prediction and to compare for the result generated from

these approaches. The main purpose of the prediction done is to let user to give correct focus and solve in the dengue outbreak issue. In this work, the SAS@ software will be used to develop the logistic regression models. Figure 2 depicted a snapshot of SAS Enterprise Miner Workstation 13.

Here, several models are developed. The best model will be selected based on the best fitting and most parsimonious and reasonable model to describe the relationship between the target variable and input variables. After estimating the coefficients and fitting the model, the next task is to conduct an assessment of the significance of the variables in the model. This will determine whether the input variables in the model are 'significantly' related to the class variable.

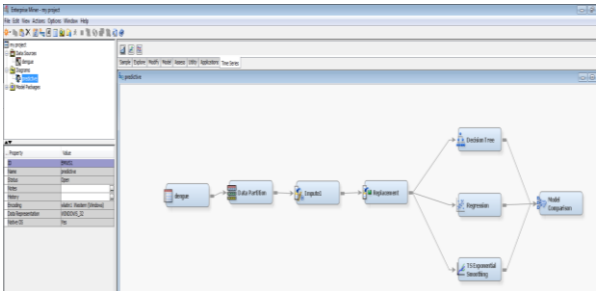
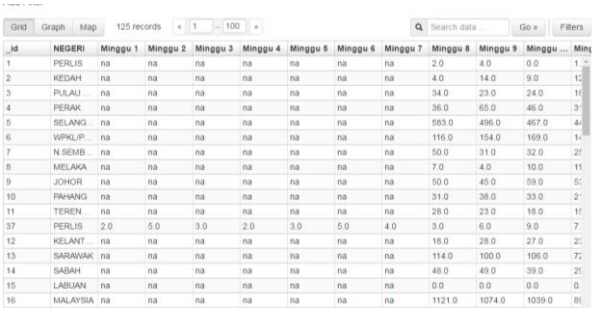


Figure 2: A Snapshot of SAS Enterprise Miner Workstation 13

IV. Result and Conclusion

The experimental result of the develop data mining application for predicting dengue fever and a dashboard system for visualizing dengue trends in Malaysia are depicted in Figure 3 to Figure 8.

A snapshot of data obtained from open.data.gov is illustrated in Figure 3. The data were arranged in the relational data format of rows and columns. Each column is designated for attributes with their values, while the final column contains the class attributes with a set of possible class labels. Each row is reserved for the items and represents one record often referred to as an 'instance'.



_id	NEGERI	Minggu 1	Minggu 2	Minggu 3	Minggu 4	Minggu 5	Minggu 6	Minggu 7	Minggu 8	Minggu 9	Minggu ...
1	PERLIS	na	na	na	na	na	na	na	2.0	4.0	0.0
2	KEDAH	na	na	na	na	na	na	na	4.0	14.0	9.0
3	PULAU ...	na	na	na	na	na	na	na	34.0	23.0	24.0
4	PERAK	na	na	na	na	na	na	na	36.0	65.0	46.0
5	SELANG...	na	na	na	na	na	na	na	583.0	496.0	467.0
6	WPKL/P...	na	na	na	na	na	na	na	116.0	154.0	169.0
7	N.SEMB...	na	na	na	na	na	na	na	50.0	31.0	32.0
8	MELAKA	na	na	na	na	na	na	na	7.0	4.0	10.0
9	JOHOR	na	na	na	na	na	na	na	50.0	45.0	59.0
10	PAHANG	na	na	na	na	na	na	na	31.0	38.0	33.0
11	TEREN...	na	na	na	na	na	na	na	28.0	23.0	18.0
37	PERLIS	2.0	5.0	3.0	2.0	3.0	5.0	4.0	3.0	6.0	9.0
12	KELANT...	na	na	na	na	na	na	na	18.0	28.0	27.0
13	SARAWAK	na	na	na	na	na	na	na	114.0	100.0	106.0
14	SABAH	na	na	na	na	na	na	na	48.0	49.0	39.0
15	LABUAN	na	na	na	na	na	na	na	0.0	0.0	0.0
16	MALAYSIA	na	na	na	na	na	na	na	1121.0	1074.0	1039.0

Figure 3: A Snapshot of Dengue Data from open.data.gov

Figure 3, shows a snapshot of the Dashboard System for Monitoring of Malaysian Dengue Trends. Based on the developed system, users are capable to visualize the information into three main functions. The first function is for accessing the actual data. The second and third menu is for predicting modeling result based on three main techniques namely the exponential smoothing, moving average and simple linear approach.

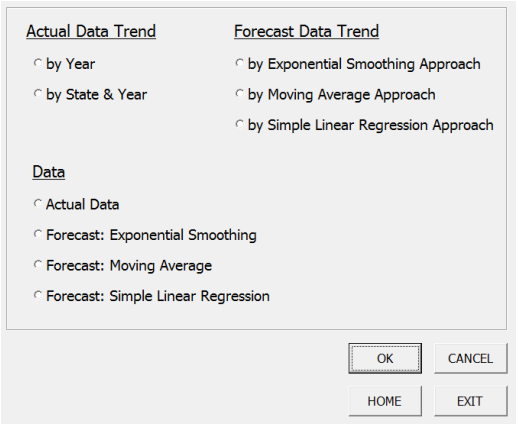


Figure 3: A Snapshot of Dashboard System for Monitoring of Malaysian Dengue Trends

The chart showed in Figure 4 summarized and visualized the whole information of dengue cases from open.data.gov. Within the period of 2010 until 2015, the highest recorded dengue cases happen in year 2014 with 5,546 cases. The chart also showed that the dengue cases will reduce drastically start from the year of 2015.

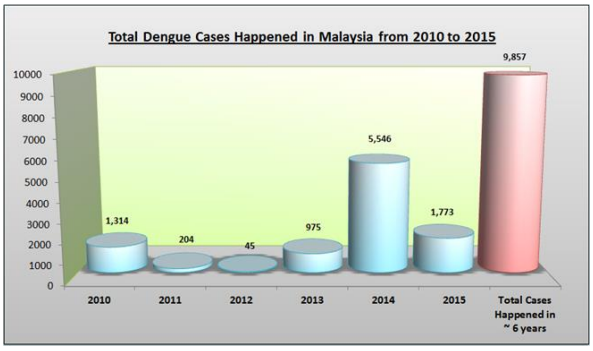


Figure 4: The Total Dengue Case Happened in Malaysia from Year 2010 to 2015

Refer to the details for the year 2014 (as Figure 5), Selangor state had the highest dengue cased happened (73.5% from overall 5,546 cases). This system could let user easier to plan some strategies to combat down the dengue outbreak issue.

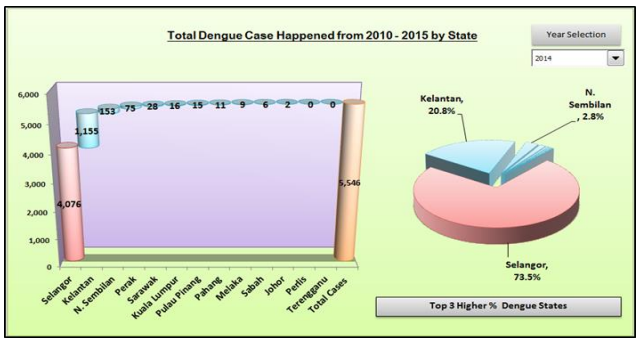


Figure 5: Dengue Case Comparison by years and by states

User also could conduct forecast analysis based on the prediction tools in this system where Figure 6 shows that the forecast done by using Exponential Smoothing approach, Figure 7 shows that the forecast done by using Moving Average approach and Figure 8 shows that the forecast done by using Simple Linear Regression approach for the next 5 years.

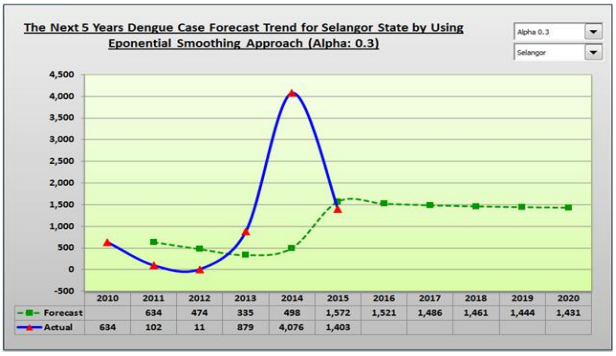


Figure 6: Forecast Dengue Case Trends of Selangor State for the Next 5 Years (by Exponential Smoothing approach)

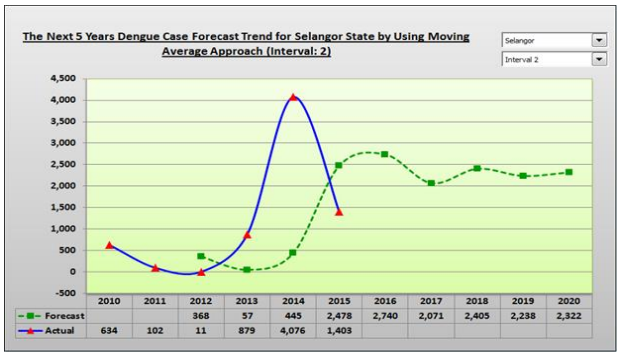


Figure 7: Forecast Dengue Case Trends of Selangor State for the Next 5 Years (by Moving Average approach)

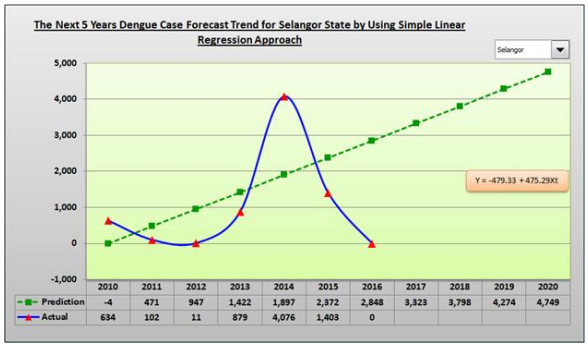


Figure 8: Forecast Dengue Case Trends of Selangor State for the Next 5 Years (by Simple Linear Regression approach)

As a conclusion, the system could easily let user to have more precise target and efficient solution to combat the dengue outbreak issue. The system also could provide the answer for the research questions where (1) The outlook of dengue case happened in Malaysia initially increase drastically from year 2010 to 2014 but the cases are reduced back start from the year 2015. (2) Selangor is the state has the top dengue cases happened (73.5%) at year 2014 which it need to give focus for prevent the outbreak. (3) Overall forecast results (Exponential Smoothing, Moving Average and Simple Linear Regression approaches) shows that the dengue outbreak trend at Malaysia will increase for the next 5 years. (4) Selangor is the state forecasted to have the highest dengue outbreak cases happen for the next 5 years.

V. Result and Conclusion

As a conclusion, the system could easily let user to have more precise target and efficient solution to combat the dengue outbreak issue. The system also could provide the answer for the research questions where (1) The outlook of dengue case happened in Malaysia initially increase drastically from year 2010 to 2014 but the cases are reduced back start from the year 2015. (2) Selangor is the state has the top dengue cases happened (73.5%) at year 2014 which it need to give focus for prevent the outbreak. (3) Overall forecast results (Exponential Smoothing, Moving Average and Simple Linear Regression approaches) shows that the dengue outbreak trend at Malaysia will increase for the next 5 years. (4) Selangor is the state forecasted to have the highest dengue outbreak cases happen for the next 5 years.

Acknowledgment

This work was supported by MOE under Fundamental Research Grant Schema.

References

[1] L. K. Ghee, A Review of Disease in Malaysia. Pelanduk Publication., 1993.
[2] D. J. Gubler, "Dengue and Dengue Hemorrhagic Fever," Clin. Microbiol. Rev., vol. 11, no. 3, pp. 480–496, Jul. 1998.
[3] K. J. McConnell and D. J. Gubler, "Guidelines on the cost-effectiveness of larval control programs to reduce dengue transmission in Puerto Rico," Rev. Panam. salud pública, vol. 14, no. 1, pp. 9–16, 2003.
[4] N. N. A. Husin, N. Salim, and A. R. A. Ahmad, "Simulation of Dengue Outbreak Prediction," Proc. Postgrad. Annu. Res. Semin., pp. 374–379, 2006.
[5] A. Munasinghe, H. Premaratne, and M. Fernando, "Towards an Early Warning System to Combat Dengue," Int. J. Comput. Sci. Electron. Eng., vol. 1, no. 2, 2013.
[6] S. Promprou, M. Jaroensutasinee, and K. Jaroensutasinee, "Forecasting dengue haemorrhagic fever cases in Southern Thailand using ARIMA Models," Dengue Bull., vol. 30, p. 99, 2006.
[7] D. Gerardi and L. Monteiro, "System identification and prediction of dengue fever incidence in Rio de Janeiro," Math. Probl. Eng., vol. 2011, pp. 1–13, 2011.
[8] M. Gharbi, P. Quenel, J. Gustave, S. Cassadou, G. L. Ruche, L. Girdary, and L. Marrama, "Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors," BMC Infect. Dis., vol. 11, no. 1, p. 166, 2011.