

**FAKULTI TEKNOLOGI MAKLUMAT DAN KOMUNIKASI  
(FTMK)**

**BITI 2513: INTRODUCTION TO DATA SCIENCE**

**REPORT**

**PREPARED BY:**

FARIS HANNAN BIN ISMAIL (B031820017)

NUR AFIQAH BINTI IBRAHIM (B031810125)

NUR AZWA BINTI ZUBIR (B031810102)

**LECTURER:**PM DR SHARIFAH SAKINAH BINTI SYED AHMAD

**Abstract**

COVID-19 epidemic in Malaysia started as a small wave of 22 cases in January 2020 through imported cases. It was followed by a bigger wave mainly from local transmissions resulting in 651 cases. The following wave saw unexpectedly three-digit number of daily cases following a mass gathering urged the government to choose a more stringent measure. A limited lock-down approach called Movement Control Order (MCO) was immediately initiated to the whole country as a way to suppress the epidemic trajectory. The lock-down causes a major socio-economic disruption thus the ability to forecast the infection dynamic is urgently required to assist the government on timely decisions. Limited testing capacity and limited epidemiological data complicate the understanding of the future infection dynamic of the COVID-19 epidemic. Three different epidemic forecasting models was used to generate forecasts of COVID-19 cases in Malaysia using daily reported cumulative case data up until 1st April 2020 from the Malaysia Ministry of Health. The forecasts were generated using a Curve Fitting Model with Probability Density Function and Skewness Effect, the SIR Model, and a System Dynamic Model. Method one based on curve fitting with probability density function estimated that the peak will be on 19th April 2020 with an estimation of 5,637 infected persons. Method two based on SIR Model estimated that the peak will be on 20th - 31st May 2020 if Movement Control (MCO) is in place with an estimation of 630,000 to 800,000 infected persons. Method three based on System Dynamic Model estimated that the peak will be on 17th May 2020 with an estimation of 22,421 infected persons. Forecasts from each of model suggested the epidemic may peak between middle of April to end of May 2020.

**Introduction**

A novel coronavirus infectious disease (COVID-19) which is caused by SARS-CoV-2 has been announced by the World Health Organization as a fatal global pandemic. The epidemic of COVID- 19 started explosively in Wuhan and spread throughout China. Mediated via a massive aviation industry, it turned into a pandemic in just two months. As of April 4, 2020, the number of cases climbed above 1 million with a death toll of over 50 000 worldwide. The global impact and the public health threat of COVID-19 is the most serious seen in a respiratory virus since the 1918 influenza pandemic. Both COVID-19 and the 1918 influenza pandemic are associated with respiratory spread, a significant percentage of infected people with asymptomatic cases transmitting infection to others, and a high fatality rate.

Globally, the epidemic curve in each country varies from exponential, uncontrolled outbreak (Italy) to slow rising, adequately controlled (Singapore). Malaysia somehow lies in the middle. As of April 5, 2020, there are 3,662 cases and 61 deaths in Malaysia. Malaysian government are taking prompt public health actions to prevent an exponential rise of cases by continuously screen and test high risk individuals, isolate patients and trace and quarantine the contacts to prevent secondary spread. These actions seemed to be adequate until a large cluster of cases occurred following a large Tabligh gathering involving more than 10,000 members in late February. The event has changed the direction of the epidemic curve in Malaysia.

Simple counts of the number of confirmed cases can be misleading indicators of the epidemic’s trajectory if these counts are limited by problems in access to care or bottlenecks in laboratory testing, or if only patients with symptoms are tested. This is where a prediction modelling may assist the authority in making decisions. Model-based predictions can help policy makers make the right decisions in a timely way, even with the uncertainties about COVID-19. We used a combination of actual daily data and analysis of patterns and trends from previous cases in other countries to predict the projection of upcoming cases for Malaysia. The projection serves to support the needs for lockdown period and activity to mitigate the spread of coronavirus cases.

**Objective**

1. To study the positive rates of COVID using the recorded data
2. Identify factors related to COVID positive rates for their respective locations
3. Contribute to product-related assessments acting as Peer Reviewer and as a forum for discussion on the rolling data assessment.

**Define goal**

1. Able to be aware of COVID dangers among the citizens.
2. Able to analyze positive and cured cases.
3. Able to reduce the risk of COVID infection.

**Data Description and Preparation**

**Dataset**

In order for the statistical and machine learning algorithms to learn and predict the trend and growth of the disease, several online news and related websites (such as the Malaysia’s official health ministry websites) was crawled and fed into a database. Covid-19 data on the number of susceptible, infectious, recovered, and deceased patients for world countries are available from Worldometer and the 2019 Novel Coronavirus Visual Dashboard. For Malaysia, daily data has been published by the STAR, Malaysiakini and also by the Ministry of Health Malaysia.

For the Malaysian COVID-19 dataset, data on medical capacity (e.g.: number of beds in each state) and events that could affect the spread of the disease (example: Tabligh Assembly at Sri Petaling Mosque) were also collected to see if this data could help in making a strategy to flatten the curve of infected cases. The world data on the dates of restriction and quarantine declared by each affected country were also collected in order to gauge and infer how the infection would pattern of COVID-19 cases in Malaysia looks like, if similar controlled measures are implemented.

**Data management**

The COVID-19 pandemic, also known as the coronavirus pandemic, is an ongoing pandemic of coronavirus disease 2019 (COVID‑19), caused by severe acute respiratory syndrome coronavirus 2 (SARS‑CoV‑2). The outbreak was first identified in Wuhan, China, in December 2019. The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 30 January, and a pandemic on 11 March. As of 9 June 2020, more than 7.12 million cases of COVID-19 have been reported in more than 188 countries and territories, resulting in more than 406,000 deaths; more than 3.29 million people have recovered.

We have extracted data for a few days until 9 June 2020, across 6 variables. This was collected for only in Malaysia. We have also extracted Data for National Security Council. This data covers new cases, cases confirmed, cases in treatment, cases discharged and deaths.



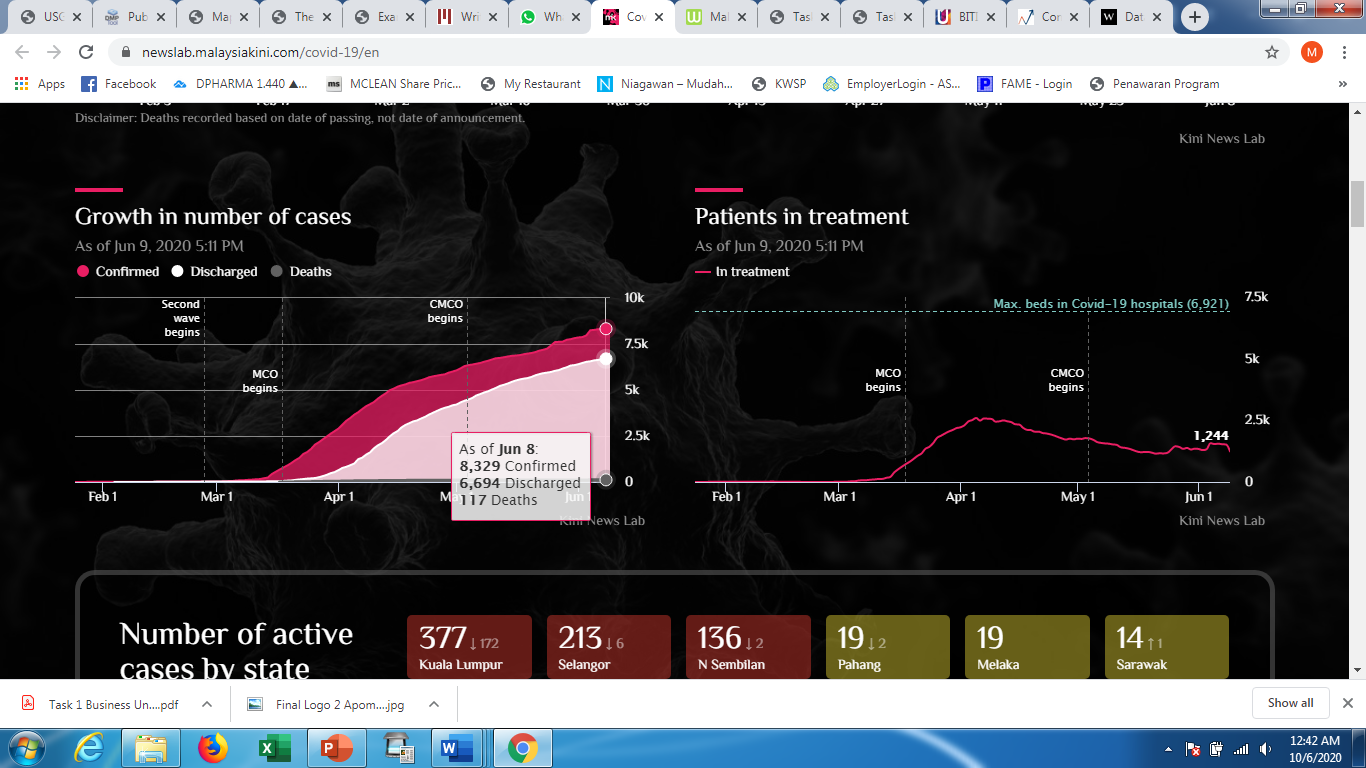
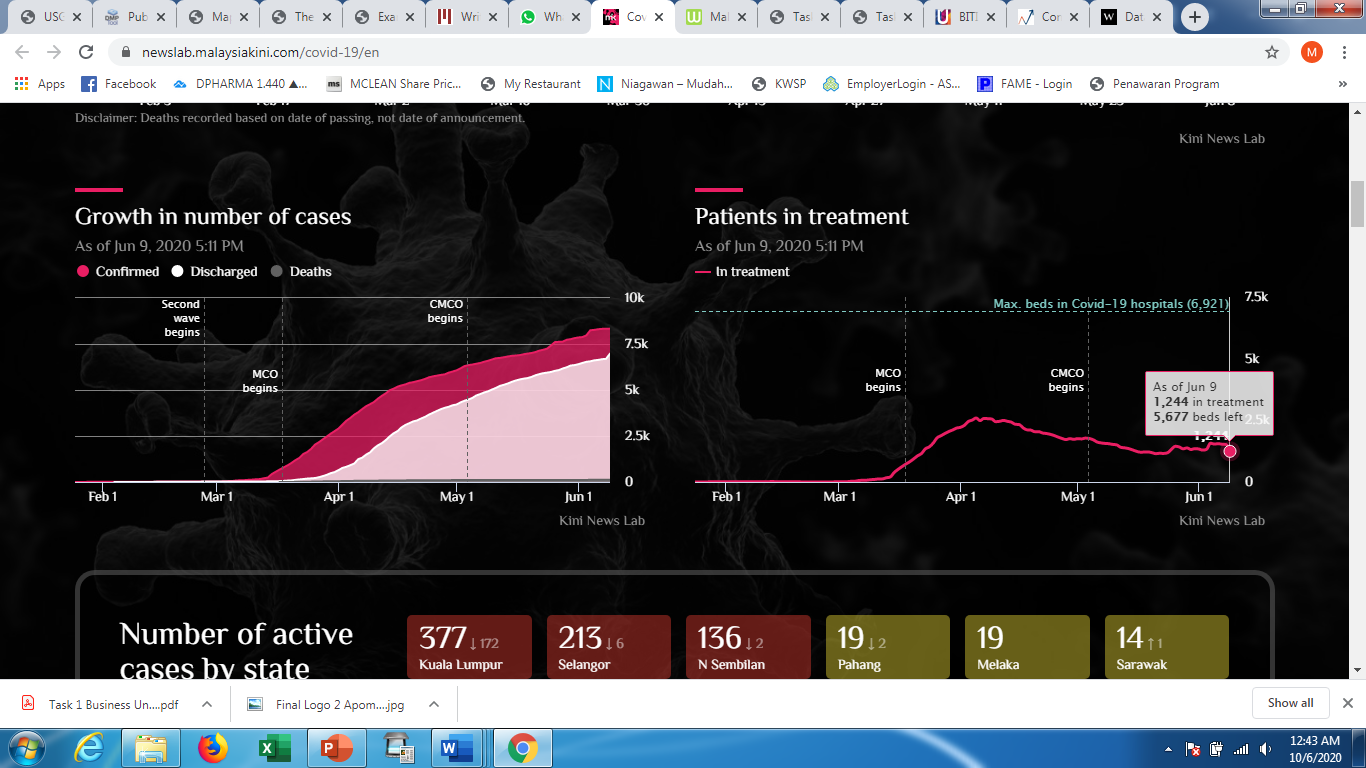
 

Table 1: Table showed list of variables included

Table 1 shows the overview of the reported confirmed cases, recovered cases, and death cases of Malaysia from 1st February 2020. It is expected that the active cases reach its peak in the middle of April. The active cases may end by end of May, on condition of full compliance of the MOC. Some uncertainties may happen, such as capacity of healthcare system, social distancing and compliance to MCO in.

Based on table, it clearly shows that few days before the implementation of the Movement Control Order (MCO) has cause drastic increase of the confirmed cases (up to 80%), and the growth rate has dropped but remaining considerably high or exponential for the rest of the days within the MCO period. This also indicate that the implementation of the MCO is effective to reduce sudden spike of the growth of the confirmed cases of COVID19. The high increment of cases from 238 to 428 cases in a day between 14th March 2020 and 15th March 2020 made a huge difference in the next twelve days. The difference is around 1250 cases, as shown in Table 1. The degree of compliance during lockdown (MCO) period is crucial.

**Data quality**

The virus is primarily spread between people during close contact, most often via small droplets produced by coughing, sneezing, and talking. Less commonly, people may become infected by touching a contaminated surface and then touching their face. Qualitative of data involved:

* New cases increase sharply during Movement Control Order (MCO) begins.
* However, it seems new cases in Conditional Movement Control Order (MCO) become decrease. Some days had data for only less than hundred for new cases.
* At the same time, looks the increasing of number of discharge.

**Data preparation**

1. **SEIR Model**

Mathematical models can be designed to stimulate the effect of disease within many levels. These models can be used to evaluate disease from within the host model i.e. influence interaction within the cells of the host to meta population model i.e. how its spread in geographically separated populations. The most important part of this model is to calculate the R0 value. The value of R0 tells about the contagiousness of disease. It is the fundamental goal of epidemiologists studying a new case. In simple terms R0 determines an average of what number of people can be affected by a single infected person over a course of time.

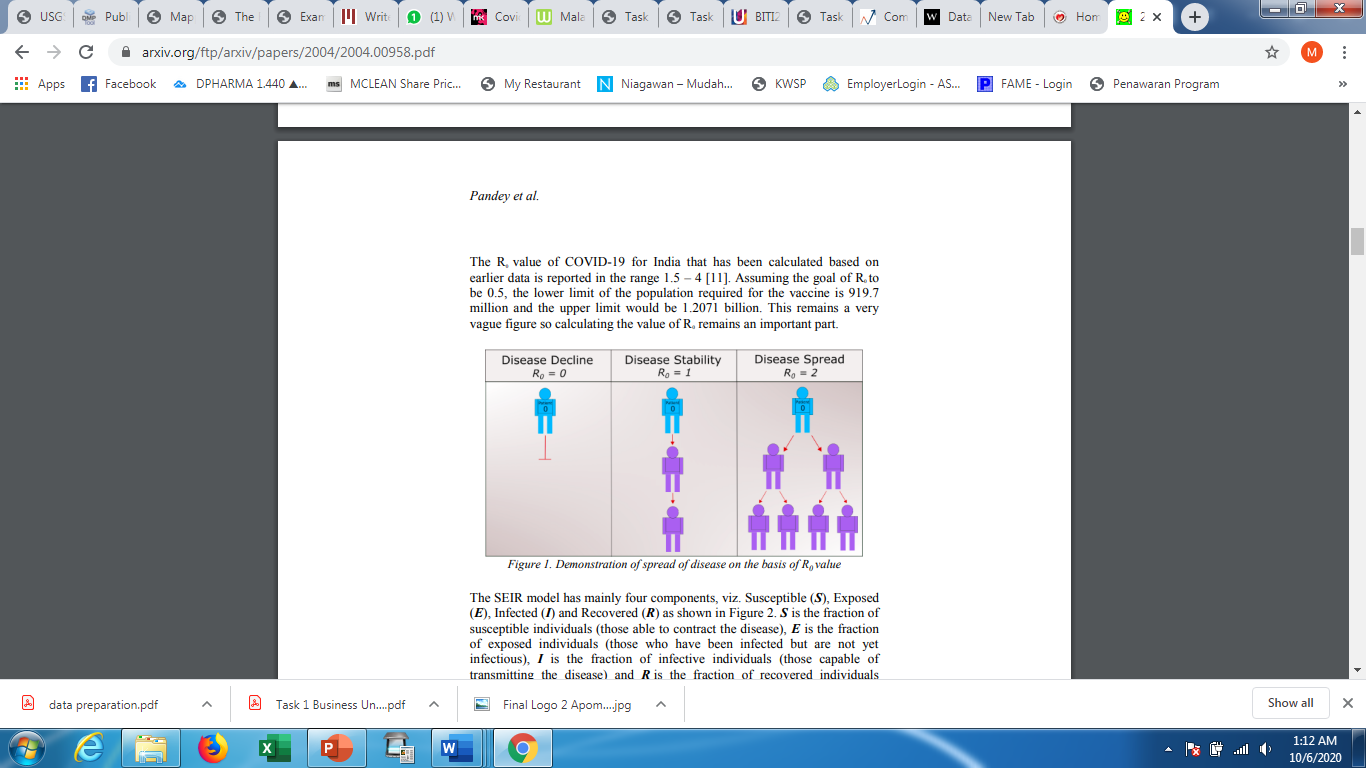


Figure 1. Demonstration of spread of disease on the basis of R0 value

1. **Regression Model**

Regression models are statistical sets of processes which are used to estimate or predict the target or dependent variable on the basis of dependent variables. The regression model has many variants like linear regression, ridge regression, stepwise regression, polynomial regression etc. This study has used linear regression and polynomial regression for prediction of COVID-19 cases. Linear regression is a simple model which is used to finds the relation between a dependent and an independent variable.

For building the Multiple Linear Regression Model, all the variables were transformed using logarithm.

**Missing values and Outliers**

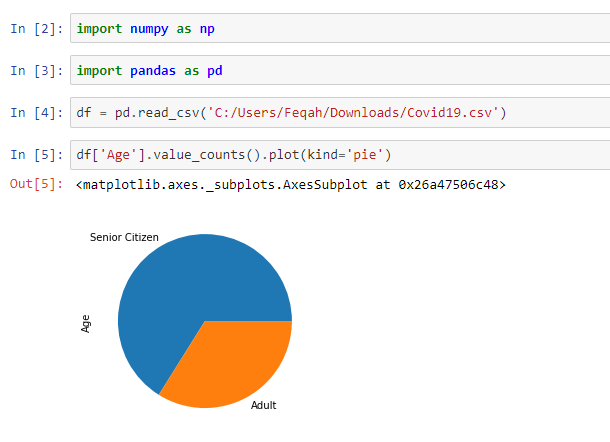
No specific missing value treatment was used. Days for which no data was available for the key variables, then that day’s record was not removed from analysis. The observations were recorded for the key variables were included in the analysis

**Exploratory Data Analysis**

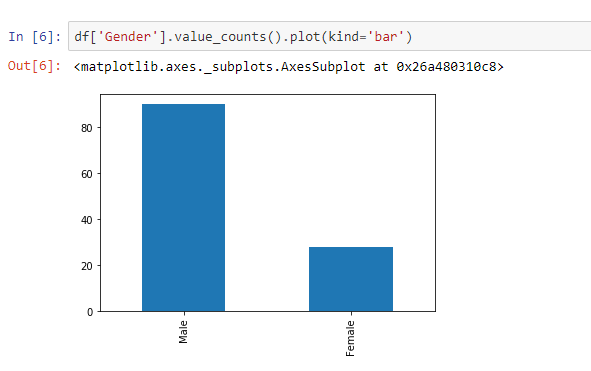
In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

The objectives of EDA are to:

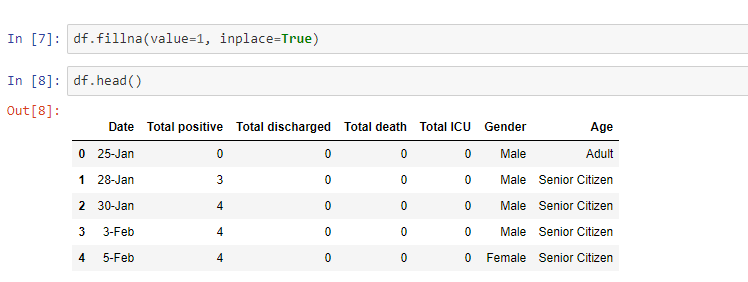
* Suggest hypotheses about the causes of observed phenomena
* Assess assumptions on which statistical inference will be based
* Support the selection of appropriate statistical tools and techniques
* Provide a basis for further data collection through surveys or experiments.



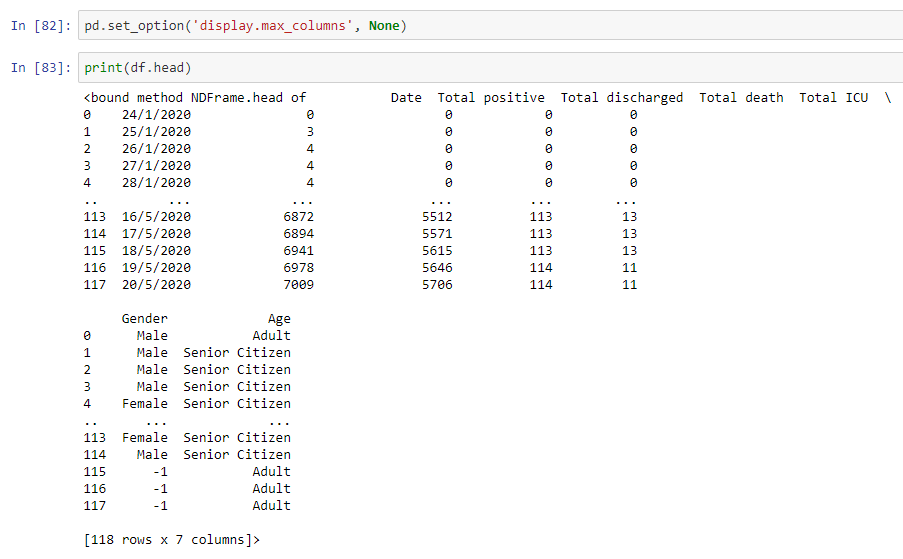
Based on the figure, we have a DataFrame with the senior citizen and adult. We pass the the ‘age’ column to the pie function to get a pie plot.

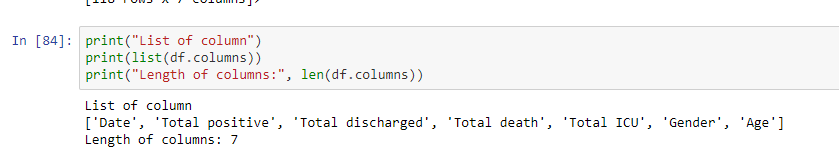


We access the gender field, call the value\_counts method to get a count of unique values, then call the plot method and pass in bar (for bar chart) to the kind argument.



Use fill() function to fill the missing values along the index axis.





**Data Analysis Plan**

The combination of visual analytics and data science enables people with little knowledge of statistics, to understand complex scenarios and draw inference about the future, from current events. The COVID-19 virus has some behavioral attributes and survival strategies that make it difficult to anticipate short and long term infection scenarios. In particular, the exponential doubling can turn an initial spark infection into a significant outbreak in a matter of weeks.

The COVID-19 virus has some behavioral attributes and survival strategies that make it difficult to anticipate short and long term infection scenarios. In particular, the exponential doubling can turn an initial spark infection into a significant outbreak in a matter of weeks.

**Outbreak Status and Update**

Figure 2 to figure 5 shows the current trends for COVID-19 outbreak as from data retrieved. The cases reported are visualized in analytics dashboard to show the outbreak trend for total positive, total death, total ICU and total discharge for all states. This aligns with our objectives to show the outbreak progress over the period of time for each segment. It was found that the total number of confirmed cases for all states increasing steadily.

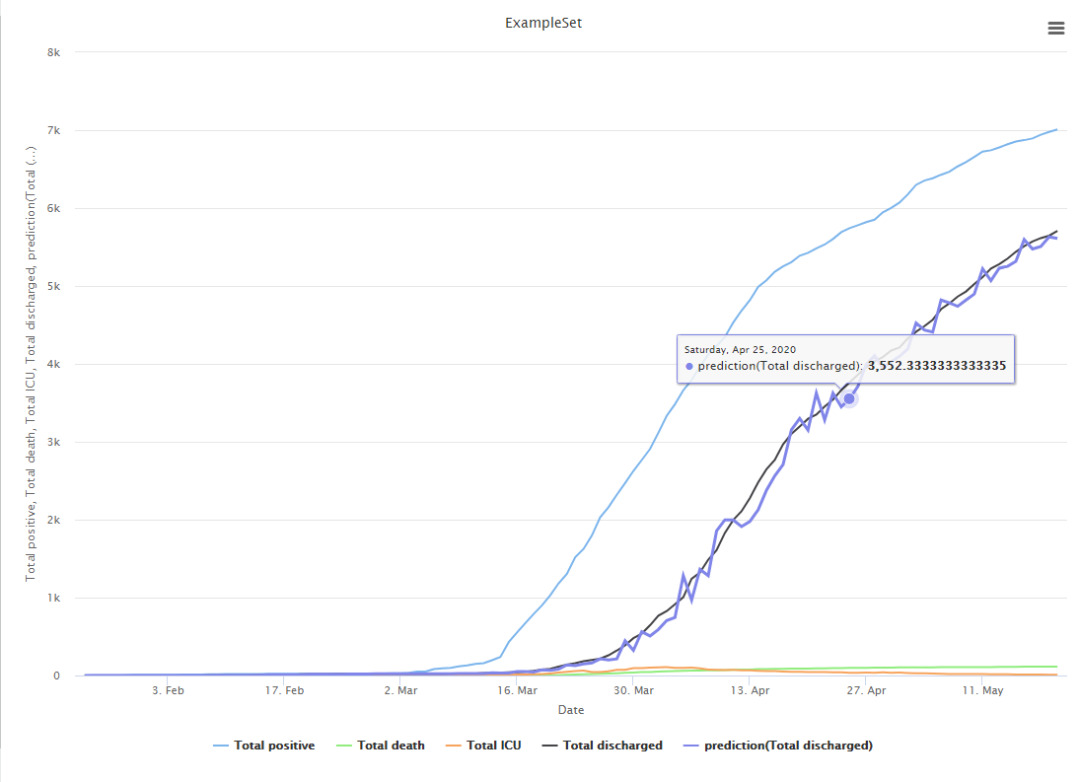


Figure 2: Decision Tree

root\_mean\_squared\_error: 105.163 +/- 28.961

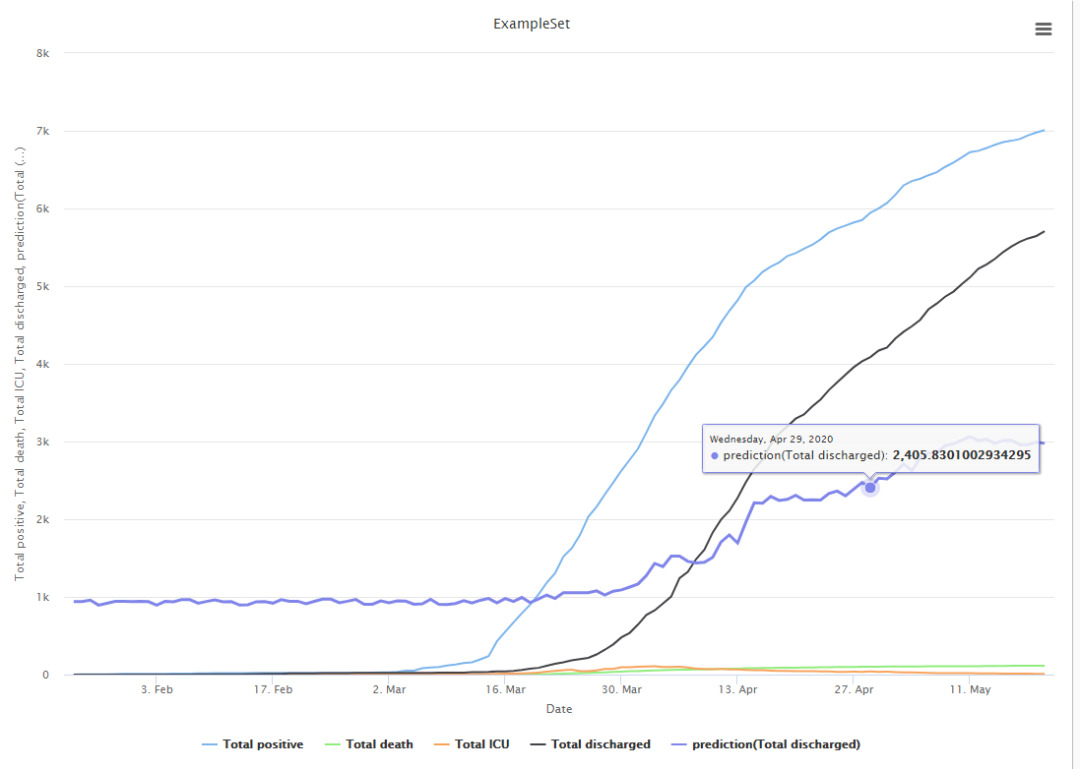


Figure 3: Gradient Decision Tree

root\_mean\_squared\_error: 1218.874 +/- 115.526

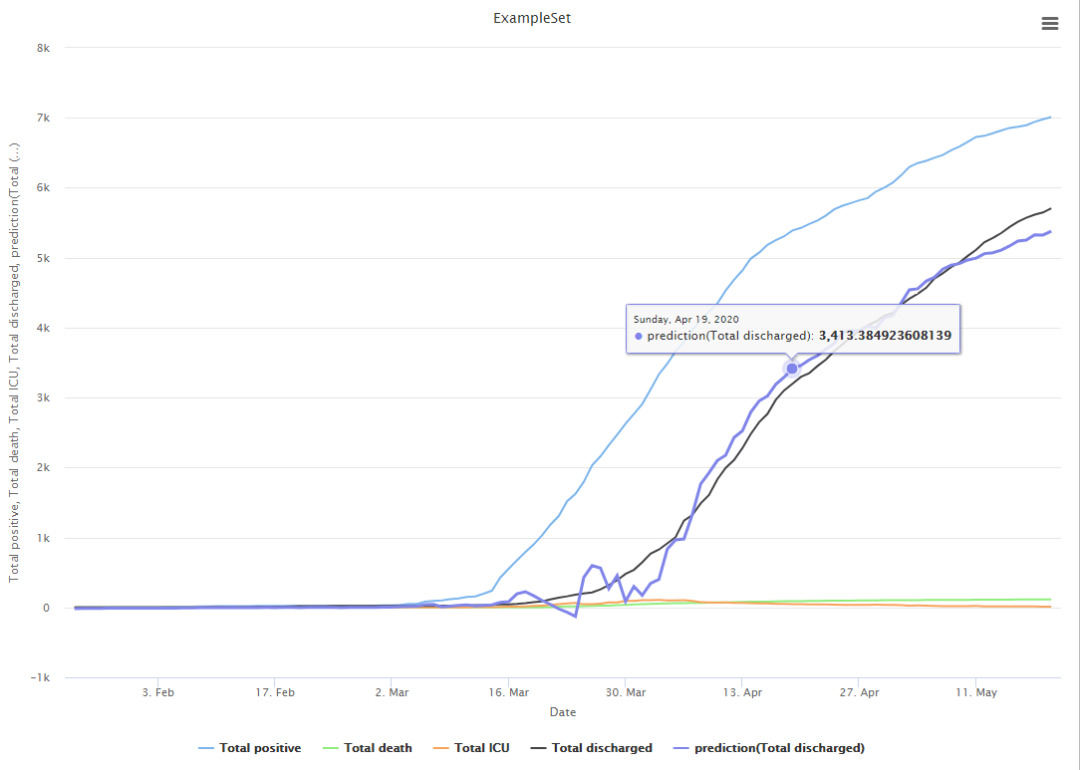


Figure 4: Linear Regression

root\_mean\_squared\_error: 156.503 +/- 47.541

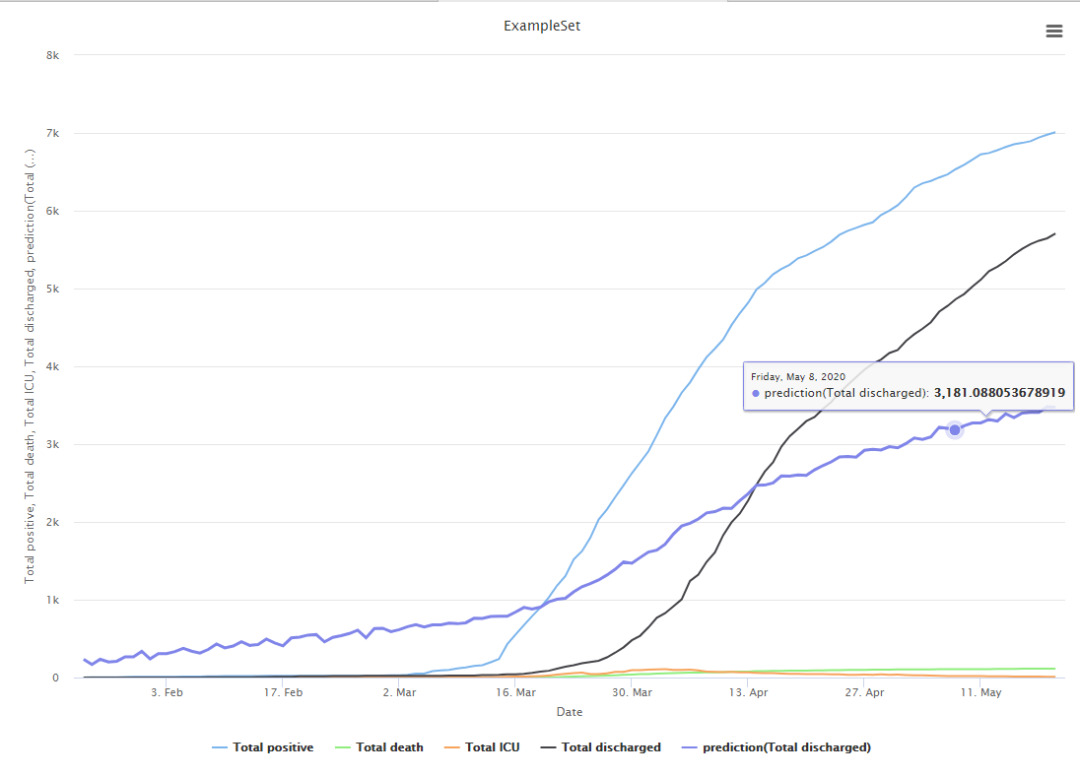


Figure 5: Generalized Linear Model

root\_mean\_squared\_error: 959.216 +/- 131.786

**Conclusion**

The reason Malaysia success on flattening the curve to prevent the overwhelming the health care system is:

1. Complete prohibition of movements and large gatherings across the country.
2. Complete restriction on all overseas travels by Malaysians; for those returning from overseas, they are required to undergo 14-day self-quarantine (but starting April 3, 2020, 14-day mandatory quarantine at a facility).
3. Complete restriction on all tourists and foreigners entering into the country.
4. Closures of kindergartens, nurseries, government and private schools.
5. Closure of all public and private higher education institutions, as well as skills training institutes.
6. Closure of all government and private premises, except those involved in providing essential services (including health).

The faster we react to the positive cases the higher the recovery rates from COVID19. We can see the data during the MCO and CMCO which significantly reduce the death rate.

**Reference**

1. <https://www.tibco.com/blog/2020/03/18/covid-19-a-visual-data-science-analysis-and-review/>
2. <https://towardsdatascience.com/exploring-covid-19-research-publications-407f8c2aa842>
3. <https://www.who.int/bulletin/online_first/20-255695.pdf>
4. <https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm>