SCHOOL OF ENGINEERING AND TECHNOLOGY

FINAL ASSESSMENT FOR THE BSC (HONS) INFORMATION SYSTEMS (BUSINESS ANALYTICS)/ (DATA ANALYTICS); YEAR 3

ACADEMIC SESSION AUGUST 2021; SEMESTER 7 and 8

BIS 3218: BUSINESS INTELLIGENCE SYSTEMS

DEADLINE: 11th DECEMBER 2021 11:59PM

GROUP NAME: OneLastRide

GROUP MEMBERS: CH'NG KE XIN (17106097 / 000403-07-0290)
CHAN WEI CHEE (16052755 / 980103-14-5218)
CHAN WEI WEI (16052748 / 980103-14-5226)
CHEW JIA HURNG (19082072 / 991116-06-5317)
LEE ZHI CHENG (17027590 / 990329-10-6149)
TEY YUEN YUEE (19000272 / 000812-14-0536)

INSTRUCTIONS TO CANDIDATES

- This project will contribute 50% to your final grade.
- This is a group project. Each group consists of 6 members.

IMPORTANT

The University requires students to adhere to submission deadlines for any form of assessment.
Penalties are applied in relation to unauthorized late submission of work.

- Coursework submitted after the deadline but within 1 week will be accepted for a maximum mark of 40%.
- Work handed in following the extension of 1 week after the original deadline will be regarded as a non-submission and marked zero.

Academic Honesty Acknowledgement

"We (names stated above) verify that this paper contains entirely my own work. I have not consulted with any outside person or materials other than what was specified (an interviewee, for example) in the assignment or the syllabus requirements. Further, I have not copied or inadvertently copied ideas, sentences, or paragraphs from another student. I realize the penalties *(refer to page 16, 5.5, Appendix 2, page 44 of the student handbook diploma and undergraduate programme)* for any kind of copying or collaboration on any assignment."

1) CH'NG KE XIN

2) CHAN WEI CHEE

3) CHAN WEI WEI

4) CHEW JIA HURNG

5) LEE ZHI CHENG

6) TEY YUEN YUEE

Ch'ng KX, Chan WC, Chan WW, Chew JH, Lee ZC, Tey YY | 11-12-2021

(Student's signature / Date)

TABLE OF CONTENTS

1. Business Understanding

1.1 Introduction

The coronavirus disease 2019 (COVID-19) is a disease caused by a virus named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which is a contagious virus under the category of coronavirus that has surfaced in late 2019, leading to the commencement of novel coronavirus [1]. The sudden emergence of COVID-19 had caused a strong hit to both the global economy and health. WHO Director General declared that the outbreak constitutes a Public Health Emergency of International Concern (PHEIC) on 30 January 2020 and a pandemic on 11 March 2020 [2] [3]. According to [4], the COVID-19 pandemic has resulted in a tremendous loss of human life globally, posing an unprecedented challenge to public health, workplace and food systems. The pandemic's economic and social impact is destructive, causing a vast amount of people are at risk of sliding into severe poverty. With the prohibition of going out during lockdowns, many were unable to earn an income in sustaining themselves and their families. Adding to it, the aviation industry is also one of the hardest hits by COVID-19. The menacing pandemic had forced an immediate suspension of global travel to avoid the flow of people to minimise the close contact between people and thus prevent transmission of the virus. The pandemic has caught the world a shock and there seem to be no signs of disappearing although it has been existing for two years as the virus is now still diffusing around the world and people are getting frustrated with normal lives affected and the loss of dear ones.

1.2 Problem Statement

The threat of COVID-19 became more obvious in Malaysia after neighbouring Singapore reported its first imported COVID-19 case from Wuhan, China on 23rd January 2020 which was also the republic's first positive case. Malaysia reported their first COVID-19-positive case on 25th January 2020 in less than 48 hours after the first case was reported in Singapore. The commencement of Movement Control Order (MCO) on 18th March 2020 had effectively reduced the active COVID-19 cases with daily confirmed cases showing a downtrend since 5th April 2020 [5]. The country thus allowed the conditional resumption of certain business operations to reduce economic losses. However, the confirmed cases had reached nearly 700 cases on the 6th October 2020 and had no sign of declining. The highest daily cases of 24,599 were reported on 26th August

2021 and the highest death cases of 592 were reported on 11th September 2021 which panic caused among Malaysians [6]. Cumulatively, 2,652,773 active COVID-19 cases and 30,614 death cases were reported as of 6th December 2021 [7]. According to [8], the pandemic fatigue experienced by Malaysians was increasingly severe and reached a concerning level. Although movement restrictions and SOP compliance were in place, the number of new daily cases were remaining in the five-figure level with no signs of abating. Many believe that pandemic fatigue may be the cause of mushrooming in COVID-19 new cases as Malaysians were becoming less attentive in adhering to SOPs that have been in place since March 2020. The ever-rising COVID-19 new and death cases in Malaysia has caught the government and public's attention, hoping that such a situation can be improved. Therefore, this work is to conduct deep analysis on covid data to assist in forecasting future outbreaks of cases with descriptive and predictive analysis on data and strategies to reduce the ever-rising of COVID-19 cases.

1.3 Objectives

COVID-19 has been the most pervasive topic since the outbreak of pandemic is gaining the event of concern and hence the goals of conducting this work are:

1. To understand the situation of COVID-19 in Malaysia.
2. To study the effectiveness of vaccine in reducing COVID-19 cases in Malaysia.
3. To study the cause for the trend of R_0 in Malaysia.

1.4 Literature Review

1.4.1 Background of COVID-19 in Malaysia

1.4.1.1 Past Pandemic

COVID-19 has been one of the terrible disease outbreaks, but not the first in history that doomed the entire civilisation. It follows the trajectory of several notorious past pandemics to engulf the life of millions. A century ago, the strain of influenza-A virus had incurred the deadliest pandemic in modern history which was known as the Spanish Flu [9], involving one-third of the global population from an estimate of 500 million South Seas and North Pole victims [10]. The outbreak of Spanish Flu started during the final months of World War I in 1918, infecting a major number of military personnel who possessed a weak immune system due to malnourishment as well as cramped and dirty conditions that had facilitated the strike of illness [11]. With the returning of troops to their respective home countries, the virus was brought together and eventually being spread to civilians worldwide. It was particularly harmful for 20 to 40 years old adults while being vulnerable to infants and older adults [11]. The pandemic was later declared over in summer 1919, with individuals who had been infected either passed away or developed immunity.

Moreover, a new strain of H1N1 pdm09 virus founded in Mexico in 2009 has brought about the rise of H1N1 Swine Flu Pandemic [10]. According to Centers of Disease Control and Prevention (CDC), there was an estimate of 60.8 million H1N1 cases detected in United States from April 2009 to April 2010, along with over 274,000 hospitalisations and nearly 12,500 death cases (mortality rate of 0.02%) due to this unique combination of influenza genes that was not previously founded in people or animals [12]. Despite possessing similar symptoms to the conventional flu viruses such as fever, cough and sore throat [13], the conventional seasonal flu vaccines were no longer useful in offering full protection to the citizens, thus urging the introduction of monovalent H1N1 pdm09 vaccines that was effective to end the pandemic in August 2010.

The severe acute respiratory syndrome (SARS) caused by SARS-CoV-1 in 2003 which regarded as the ancestor strain for SARS-CoV-2 that has resulted in the COVID-19 pandemic in late 2019. SARS was initially discovered in Guangdong, China in November 2002, later being brought to Hong Kong in February 2003 before spreading massively worldwide [14]. Despite having only around 8000 infected cases, it had caused a high mortality rate of 11% [15]. The incubation period took 4 to 6 days to exhibit symptoms such as high fever, headache and muscle pain. It was the first transmissible new disease in the 21st century that can be spread through respiratory droplets [16]. However, the pandemic soon came to an end after vaccines were invented.

1.4.1.2 COVID-19 Trend

The origin of COVID-19 outbreak was traced back to the first case arriving on Malaysian shores on 25th January 2020 when a passenger from China was tested positive for coronavirus. Since then, several waves of the outbreak have started and continued to date [17].

The reported cases in Malaysia were divided into three waves. The first case in Malaysia was detected on 25th January 2020 and traced back to travellers from China arriving via Singapore. The reported cases in Malaysia remained relatively low until several local clusters emerged in March 2020. The most notable was a religious event that took place at Sri Petaling, Kuala Lumpur which had led to a massive spike in cases. Consequently, the number of positive cases increased beyond 2000 active cases across every state in the country. In addition, other clusters were formed from local mass gatherings and imported cases of a person under investigation travelling from overseas countries. The third wave of COVID-19 infections in the country occurred as a result of the Sabah state election in September 2020. The third wave had been by far the most severe of the three waves, having more than 1000 daily new cases reported consistently since 3rd November 2020 [17].

As a mitigation strategy, the government of Malaysia had imposed Movement Control Order (MCO) by phases including Conditional Movement Control Order (CMCO), Recovery Movement Control Order (RMCO) to control the spread of COVID-19 nationwide. Besides, the government's response to the evolving outbreak had been adaptive and varied across time. Enforcement of regulations and standard operating procedures (SOP) were adjusted throughout the whole COVID-19 timeline. A series of precautionary measures were implemented during MCO to curb the outbreaks of COVID-19 in the country including the prohibition of mass gatherings, restrictions on foreigners entering the country, social distancing, quarantine and shutting down all facilities except primary and essential services such as health services, water, electricity, banking, transportation, telecommunication and food supply industry [18]. MCO was included in the four-phase National Recovery Plan (NRP) to assist the country to be free from COVID-19 pandemic and its economic fallout.

1.4.1.3 COVID-19 Symptoms

The clinical manifestations of COVID-19 are the signs to discover the infection of disease and they are susceptible to all walks of life. The trigger of respiratory tract infection by the disease can affect the upper respiratory tract (nose and throat) or lower respiratory tract (windpipe and lungs) which infections are regarded from mild to fatal [19]. According to [20], symptoms often show as fever, cough, loss of taste or smell, headache, fatigue, congestion or runny rose, shortness of breath, nausea or vomiting, diarrhoea and sore throat. The COVID-19 common features of fever, cough, fatigue and dyspnoea are affirmed under 95% confidence interval [21]. Another study by [22] also illustrated fever (83%-98%), cough (76%–82%) and dyspnoea (31%–55%) are the typical symptoms of such disease. According to [23], the average period from the onset of symptoms to hospitalisation was 7 days. However, it only requires a day more (8 days) for the disease to exacerbate to shortness of breath and the need for ventilation in approximately 39% of patients was 10.5 days. Death and adverse outcomes are more prevalent among the elderly and those with underlying comorbidities had contributed to 50-75% of fatal cases [23].

1.4.1.4 COVID-19 Vaccination

To combat COVID-19, immunisation plays a critical role in the prevention and control of infectious disease outbreaks. It is the process of becoming immune or being protected against a disease through vaccination. Vaccines stimulate human’s immune systems to protect them from COVID-19. The impact of vaccines on health, economic and social benefits was significant as discussed by [24]. Reduction in the morbidity and mortality of infectious diseases are the major health benefits of vaccination. According to [25], vaccination has significantly reduced the COVID-19 infection in Malaysia by 88% ever since Malaysia began the COVID-19 immunisation programme in February 2021.

As immune response to different infections varies, with some being weaker or fading over time, some vaccines require two doses to reinforce the immune response and maximise the protection against the disease. The increase of antibody production to a level that protects a person against COVID-19 is boosted by the second dose vaccine. Vaccination does not result in an immediate immunity against the disease, and it takes time to build up immunity. Therefore, a person is only considered fully vaccinated at least 14 days after his second dose vaccination [25]. Table 1-1 displayed the comparison of three vaccines that are commonly adopted in Malaysia [26] [27].

Table 1-1. Comparison of Vaccines

	Pfizer	**Sinovac**	**AstraZeneca**
Type	MRNA vaccine	Inactivated virus vaccine	Adenovirus vector vaccine
Minimum Age	12 years old	18 years old	18 years old
Dosage	2 doses	2 doses	2 doses
Minimum Number of Days Apart	21 days	14 days	28-84 days
Efficacy	95%	51%	70.4%
Origin Country	United States	China	United Kingdom

1.4.1.5 R Naught (R_0)

R naught (R_0) is referred as the basic reproduction number that is capable to quantify the contagiousness of infectious pathogens such as coronavirus [28]. It is commonly used to represent the average number of secondary infections stemming from a single infectious case in the susceptible community [29]. The absence of standard mathematical equation to derive such metric implies that the calculation encapsulates the infectious period, mode of transmission and the contact rate of a positive patient [30]. The computed ratio is a positive value, where it is benchmarked at $R_0 = 1$ to represent the stability of the disease outbreak in not causing a pandemic. The government recognises $R_0 < 1$ as the recommended safety level given the low transmissibility of the disease [28].

The China government had introduced travel restrictions on 25th January 2020 after discovering confirmed cases of COVID-19 in Wuhan, Hubei Province. According to [31], the implementation of such lockdown had successfully led to a rapid decline of median R_0 value from 2.35 to 1.05 in just 2 weeks. The changes of R_0 value are then used to indicate the decrease in the severity of COVID-19 in China that is achieved through the reduced contact rates among the Chinese population. Hence, this epidemiological measure is commonly utilised to suggest the urgency of healthcare and government intervention practices to mitigate the transmissibility of the pandemic [30].

1.4.2 Modelling Technique used on COVID-19 Related Work

1.4.2.1 Clustering

Clustering is a technique used to segregate groups with similar traits and assign them into groups or clusters. A cluster is described as a group of data objects that are more similar to each other than observations within other groups. Centroid-based, hierarchical-based and density-based methods are the various types of clustering.

There are several k-means cluster analyses conducted in the COVID-19 context. [32] aims to reveal the characteristics of patients due to exposure to the COVID-19 in the Indonesian Navy with the characteristics of sex, age, and comorbidities in the mortalities of COVID-19 patients of Indonesian Navy Personnel. The study concluded that deaths of COVID-19 are closely related to age, gender and comorbidities.

Besides, [33] presented k-means clustering algorithms to determine the clusters according to the health care quality of different countries during COVID-19 pandemic. The study showed that results of the k-means clustering are affected on the assignment of initial centroids.

A study from [34] aimed to cluster the countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators to study the factors closely involved in the spread of disease. The clusters of the counties depicted the aspects that lead to a higher number of COVID-19 confirmed and death cases and led to the evaluation of countries’ strategies in mitigating the pandemic.

Table 1-2. Related Work on Clustering

Paper/Author	Method/Techniques	Study Context
Clustering of Countries for COVID-19 Cases based on Disease Prevalence, Health Systems and Environmental Indicators [32]	K-Means Clustering	Characteristic of COVID-19 death patients
An Efficient K-means Clustering Algorithm for Analysing COVID-19 [33]	K-Means Clustering	Health Care Quality Clusters of Countries
K-Means Cluster Analysis of Sex, Age, and Comorbidities in the Mortalities of Covid-19 Patients of Indonesian Navy Personnel [34]	K-Means Clustering	Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators

1.4.2.2 Time Series

Time series analysis is a technique that uses statistics approach in handling data related to time. Time series data can be in a series of intervals (days/weeks) [35]. The purpose of conducting a time series analysis is to understand the nature and pattern of the time series data and forecast the upcoming value of the time series variable. According to Statistics Solutions [35], there are three types of data in time series analysis: Time Series Data, Cross-Sectional Data, and Pooled Data. Time Series Data refers to data collected is focused on a subject at different times; Cross-Sectional Data refers to a collection of data collected at the same time but focused on multiple subjects; Pooled Data is the combination of time series and cross-sectional data. In this study, the scope is on time series data as the dataset focus on a single subject (Daily COVID-19 cases in Malaysia). Moreover, there are numerous types of methods/models to forecast future value on a time series analysis such as moving average, exponential smoothing, and Autoregressive Integrated Moving Average (ARIMA).

Study conducted by [36] aimed to forecast the confirmed and recovered cases of COVID-19. The modelling method used by the author was an autoregressive model called Two-Piece scale mixture normal distributions (TP-SMN-AR). The result showed that the model accuracy (Mean Relative Percentage Error) for confirmed cases and recovered cases were 0.22% and 1.6% respectively. The author concluded that the model performed well in forecasting the confirmed cases and recovered cases globally.

Another study conducted by [37] used time series analysis to evaluate the main features in forecasting the trends and possible stopping time of the COVID-19 outbreak in Canada and globally. The author utilised Long Short-Term Memory (LSTM) network for analysis, a deep learning approach for time series analysis to predict future cases of COVID-19. The model result showed that the model has Root Mean Square error (RMSE) of 34.83 and 93.4% accuracy for short-term prediction while the RMSE and accuracy for long-term prediction were 45.70 and 92.67% respectively. The researcher concluded that the pandemic was expected to end within three months (August 2020) and few infection clusters can occur until December 2020.

Furthermore, Professor Dothang Truong posted an article on LinkedIn about the forecasting model on new COVID-19 cases [38]. The data in this study was scraped from worldmeters website, and Professor Dothang used SAS studio to build the forecast model with ARIMA time series method. The model showed that new cases in the United States on 7th April 2020 could be as low as 18,000 new cases or to the worst 50,000 new cases. The author remarked that the time-series analysis is a convenient and robust tool for predicting new daily COVID-19 cases [38].

Table 1-3. Related Work on Time Series Analysis

Paper/Author	Method/Techniques	Study Context
Time Series Modelling to Forecast the Confirmed and Recovered Cases of COVID-19 [36]	Time Series Analysis Two-Piece scale mixture normal distributions (TP-SMN-AR)	Predicting future COVID-19 cases
Time Series Forecasting of COVID-19 Transmission in Canada using LSTM Networks [37]	Time Series Analysis Long short-term memory network (A deep learning approach)	
COVID-19 New Cases Forecasting Model [38]	Autoregressive Integrated Moving Average (ARIMA), SAS Studio	

1.4.2.3 Decision Tree

Decision Tree is one of the most popular predictive modelling techniques used in machine learning. It is easy-to-understand and is widely used to visually represent the decisions for different problems [39]. A decision tree generally can cover regression and classification problems and depending on the problems that the study is trying to solve, if the problem is to predict either of the two possible outcomes such as ‘Yes’ or ‘No’, the model would be known as a classification tree, and if the predicted outcome is continuous such as numbers, the model would be a regression tree [40].

There are several studies related to the context of COVID-19 done with the use of Decision Tree. For instance, a study [41] explored the growth of COVID-19 and to predict the signs of early infection containment in different countries with the use of decision tree. The study suggested that there are three factors used to predict, which are percentage of days being in lockdown, the number of days since the start of lockdown, and the death rate per million populations. With these factors, the model had successfully produced an accuracy of 80.95% in predicting early containment. On the other hand, [42] used decision tree in the study to predict the number of COVID-19 cases in different types of lockdowns implemented, and it is found that decision tree performed well in forecasting the number of confirmed cases and death cases in the next 10 days. In addition, a study [43] had also been done to predict the seriousness of COVID-19 patients with decision tree, and it had produced the highest accuracy of 94.5%. As a whole, the model showed that whether if the patients have contact with another COVID-19 patient is the most significant feature of determining if the patient is COVID-19 compromised.

Table 1-4. Related Work on Decision Tree

Paper/Author	Method/Techniques	Study Context
Exploring the growth of COVID-19 cases using exponential modelling across 42 countries and predicting signs of early containment using machine learning [41]	Exponential Growth Model, Logistic Regression, Decision Tree, Random Forest, SVM	Exploring Growth of Covid-19 and Predict Signs of Early Containment
Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types [42]	Random Forest, Decision Tree, Polynomial Regression, KNN, SVM, Time Series	Predict Covid-19 cases in different lockdown types
Prediction of COVID-19 Patient using Supervised Machine Learning Algorithm [43]	Nave Bayes, SVM, Logistic Regression, K-Nearest Neighbour (KNN), Decision Tree	Predict Covid-19 seriousness in patients

2. Data Understanding

In this study, the datasets are obtained from 2 sources, which are the Github repository of Ministry of Health Malaysia (MoH) [44] and MoH's COVID-19 official website [45]. Since the availability of vaccination program in Malaysia to the adult population started from May 2021 onwards [46], the analysis is therefore conducted using data ranging from 1st May 2021 to 31st October 2021 only to yield a better focus in meeting the objectives of the study. SAS Enterprise Guide was used to conduct pre-processing and descriptive statistics, followed by SAS Enterprise Miner for predictive analysis. The field of interest in various selected datasets along with their respective justifications are tabulated in Table 2-1.

Table 2-1. Data Understanding and its Relation to the Study Context

Objective	Problem description	Dataset	Selected fields		Justification
To understand the situation of COVID-19 in Malaysia.	What is the distribution of COVID-19 cases in each state?	cases_state.csv	date	Date (data update as of 1200hrs)	To examine the total COVID-19 cases by state.
			state	Name of state	
			cases_new	New cases reported in 24 hours window	
	What is the distribution of COVID-19 death cases in each state?	deaths_state.csv	date	Date (data update as of 1200hrs)	To examine the total COVID-19 death cases by state.
			state	Name of state	
			deaths_new	New COVID-19 death cases reported to public	
	What is the COVID-19 death rate by month?	deaths_malaysia.csv	date	Date (data update as of 1200hrs)	To understand Malaysia's overall COVID-19 death rate by month.
			deaths_new	New COVID-19 death cases reported to public	
		cases_malaysia.csv	cases_new	New cases reported in 24 hours window	
	What are the characteristics of COVID-19 death patients?	linelist_deaths.csv	date	Date of death	To segment COVID-19 death patients into clusters based
			date_positive	Date of positive sample	
			date_dose1	Date of individual's 1st dose vaccine	

			date_dose2	Date of individual’s 2nd dose vaccine	on their characteristics.
			brand1	Brand of 1st dose vaccine	
			brand2	Brand of 2nd dose vaccine	
			state	State of residence	
			age	Age (rounded to the nearest integer)	
			male	Gender (1=male, 0=female)	
			bid	1=brought-in dead, 0=inpatient dead	
			malaysian	Nationality (1=Malaysian, 0=non-Malaysian)	
			comorb	Comorbidities (1=has comorbidities, 0=no comorbidities)	
	What is the age distribution of COVID-19 death cases?	linelist_deaths.csv	age	Age (rounded to the nearest integer)	To illustrate the age distribution of COVID-19 death cases for different age categories.
	What is the daily COVID-19 confirmed cases in November 2021 in Malaysia?	cases_malaysia.csv	date	Date (data update as of 1200hrs)	To visualise the trend of COVID-19 cases over a period.
			cases_new	New cases reported in 24 hours window	
To study the effectiveness of vaccine in reducing COVID-19 cases in Malaysia.	What are the total COVID-19 confirmed cases by vaccination status?	cases_malaysia.csv	date	Date (data update as of 1200hrs)	To understand the total number of COVID-19 confirmed cases by vaccination status.
			cases_unvax	Number of non-vaccinated individuals who tested positive	
			cases_pvax	Number of partially vaccinated individuals who tested positive	
			cases_fvax	Number of fully vaccinated individuals who tested positive	
	What are the total COVID-19 death cases by vaccination status?	deaths_malaysia.csv	date	Date (data update as of 1200hrs)	To understand the total number of COVID-19 death
			deaths_unvax	Number of non-vaccinated individuals who died	

			deaths_pvax	Number of partially vaccinated individuals who died	cases by vaccination status.
			deaths_fvax	Number of fully vaccinated individuals who died	
	What is the trend of COVID-19 cases when more populations are fully vaccinated?	vax_malaysia.csv	date	Date (data update as of 2359hrs)	To compare the trend of daily fully vaccinated numbers and cumulative percentage of vaccinated populations on the trend COVID-19 new cases.
			daily_full	Number of 2nd dose vaccine administration	
			cumul_full	Cumulative frequency of individuals who completed their vaccination regimens	
		cases_malaysia.csv	cases_new	New cases reported in 24 hours window	
	How many percent of the populations are fully vaccinated when $R_0 < 1$?	cases_malaysia.csv	date	Date (data update as of 1200hrs)	To visualise the trend of R_0 values on top of the cumulative percentage of fully vaccinated populations.
			cases_new	New cases reported in 24 hours window	
		rnaught_malaysia.csv	rnaught	Daily R Naught value	
		vax_malaysia.csv	cumul_full	Cumulative frequency of individuals who completed their vaccination regimens	
To study the cause for the trend of R_0 in Malaysia	What are the factors affecting daily R_0 values?	cases_malaysia.csv	date	Date (date update as of 1200hrs)	To produce a Decision Tree model that determines the factors affecting reproduction number of Covid-19 (R_0).
			cases_new	New cases reported in 24 hours window	
			cases_recovered	New Covid-19 recovered cases reported to public	
		deaths_malaysia.csv	deaths_new	New COVID-19 death cases reported to public	
		vax_malaysia.csv	daily_partial	Number of 1st dose vaccine administration	
			daily_full	Number of 2nd dose vaccine administration	
		rnaught_malaysia.csv	rnaught	Daily R Naught value	

3. Data Preparation

Data preparations are required to be conducted on datasets which attributes cannot be directly deployed. As such, several variables are derived from the existing fields for better structure of the analysis. The linelist_deaths dataset contains columns depicting the demographic information of COVID-19 death patients. As the first objective is interested to understand the situation of COVID-19 in Malaysia as well as clustering COVID-19 death patients based on their characteristics, a new column was created to compute the vaccination status of the COVID-19 death patients. Firstly, the number of days since Dose 1/Dose 2 was calculated by subtracting the date of dose 1 or dose 2 taken from the date of COVID-19 death patients being announced positive. Following, the vaccination status of COVID-19 death patients was derived based on the number of days since dose 1 and dose 2 were taken as shown in Table 3-1.

Table 3-1. Formula of Derived Variable

Derived Variables	Formula
Day_Dose1	date_of_positive – date_dose1
Day_Dose2	date_of_positive – date_dose2

Vaccination_Status	death_nvax = Day_Dose1 < 1
	death_fvax = Day_Dose1 > 1 AND Day_Dose2 > 14
	death_pvax = Day_Dose1 > 1 AND Day_Dose2 < 14

The vax_malaysia dataset that is useful to understand the effectiveness of vaccines, contains the date of 2nd dose vaccination. However, it cannot be used to represent the full vaccination date as an individual will only be considered fully vaccinated 14 days after receiving the 2nd dose vaccine [25]. Therefore, 14 was added to the aforementioned date to derive the date that an individual was fully vaccinated, labelled as fvax_date. Moreover, the cumulative number of fully vaccinated in this dataset was used to derive the cumulative percentage of the fully vaccinated population in Malaysia by dividing the said column on 32,657,400 that represents the total Malaysian populations. All the datasets were then merged accordingly to cater the requirements of the analysis, then truncated to include only observations from 1st May 2021 to 31st October 2021 for a better focus

4. Modelling

4.1 Clustering

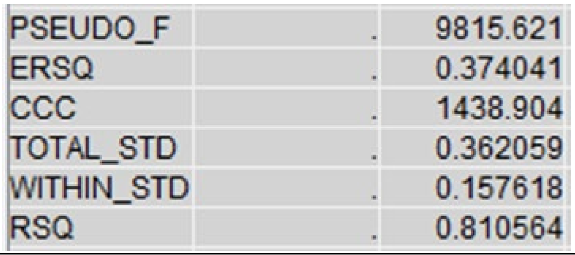
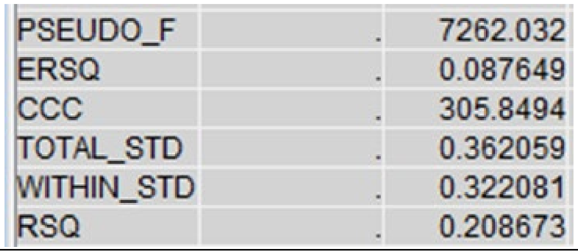
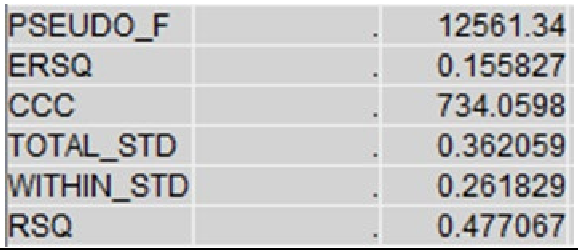
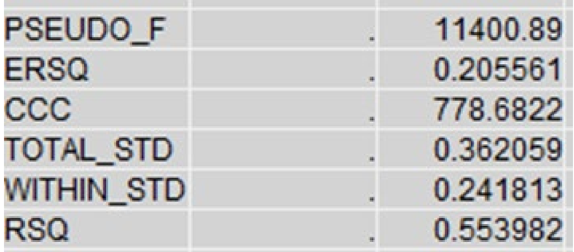
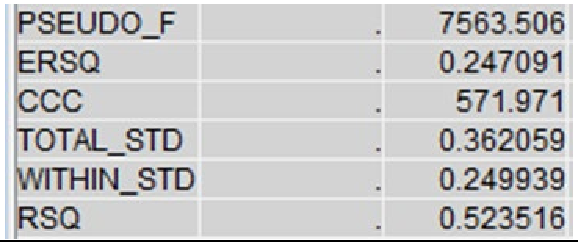
The clustering analysis was carried out using SAS Enterprise Miner. Firstly, the pre-processed dataset 'linelist_deaths_new' that consists of 19 variables was imported using file import nodes. Only 6 related variables were chosen as input parameters which are age, brand, comorb, malaysian, male and vaccination_status, with their corresponding descriptions listed in Table 4-1.

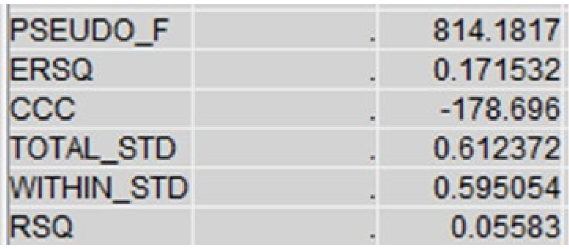
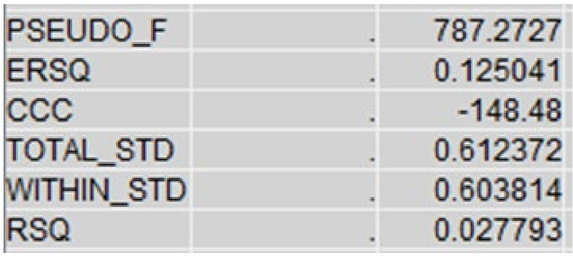
Table 4-1. Attributes used for Clustering Analysis

Attributes	Description
age	Age as an integer
brand	Brand of Vaccine
comorb	Individuals have comorbidities or do not have comorbidities
malaysian	Nationality of COVID-19 death patients
male	Gender of COVID-19 death patients
vaccination status	Vaccination status of COVID-19 death patients

The data distribution has been examined through the Stat Explore node to make sure the data points in the variables fall within the distribution range and are not skewed because clustering analysis is sensitive to the outliers. Next, the Cluster and Segment Profile node was linked from the file import node. As shown in Table 4-2, different parameter settings on the cluster node have been experimented with to find the optimal performance of clustering based on statistical measurements such as Cubic Clustering Criterion (CCC), R-Square and Pseudo-F. The diagram flow for the clustering analysis is shown in Figure 4-1.

Table 4-2. Parameter Settings and the corresponding CCC, R-Square and Pseudo-F Result

ID	Clustering Parameter Settings	Screenshot of CCC, R-Square, Pseudo-F
1	**Centroid – 10 clusters** Specification Method: Automatic Preliminary Maximum: 50 Minimum: 2 Final Maximum: 20 CCC Cutoff: 3	 PSEUDO_F . 9815.621 ERSQ . 0.374041 CCC . 1438.904 TOTAL_STD . 0.362059 WITHIN_STD . 0.157618 RSQ . 0.810564
2	**Centroid - 2 clusters** Specification Method: User Specify Maximum Number: 2 Preliminary Maximum: 50 Minimum: 2 Final Maximum: 20 CCC Cutoff: 3	 PSEUDO_F . 7262.032 ERSQ . 0.087649 CCC . 305.8494 TOTAL_STD . 0.362059 WITHIN_STD . 0.322081 RSQ . 0.208673
3	**Centroid - 3 clusters [BEST]** Specification Method: User Specify Maximum Number: 3 Preliminary Maximum: 50 Minimum: 2 Final Maximum: 20 CCC Cutoff: 3	 PSEUDO_F . 12561.34 ERSQ . 0.155827 CCC . 734.0598 TOTAL_STD . 0.362059 WITHIN_STD . 0.261829 RSQ . 0.477067
4	**Centroid – 4 clusters [2ND BEST]** Specification Method: User Specify Maximum Number: 4 Preliminary Maximum: 50 Minimum: 2 Final Maximum: 20 CCC Cutoff: 3	 PSEUDO_F . 11400.89 ERSQ . 0.205561 CCC . 778.6822 TOTAL_STD . 0.362059 WITHIN_STD . 0.241813 RSQ . 0.553982
5	**Centroid – 5 clusters** Specification Method: User Specify Maximum Number: 5 Preliminary Maximum: 50 Minimum: 2 Final Maximum: 20 CCC Cutoff: 3	 PSEUDO_F . 7563.506 ERSQ . 0.247091 CCC . 571.971 TOTAL_STD . 0.362059 WITHIN_STD . 0.249939 RSQ . 0.523516

6	**Ward - 3 clusters** Specification Method: Automatic Maximum Number: 10 Preliminary Maximum: 50 Minimum: 2 Final Maximum: 20 CCC Cutoff: 3	 PSEUDO_F . 814.1817 ERSQ . 0.171532 CCC . -178.696 TOTAL_STD . 0.612372 WITHIN_STD . 0.595054 RSQ . 0.05583
7	**Ward - 6 clusters** Specification Method: Automatic Maximum Number: 10 Preliminary Maximum: 6 Minimum: 3 Final Maximum: 6 CCC Cutoff: 3	 PSEUDO_F . 787.2727 ERSQ . 0.125041 CCC . -148.48 TOTAL_STD . 0.612372 WITHIN_STD . 0.603814 RSQ . 0.027793

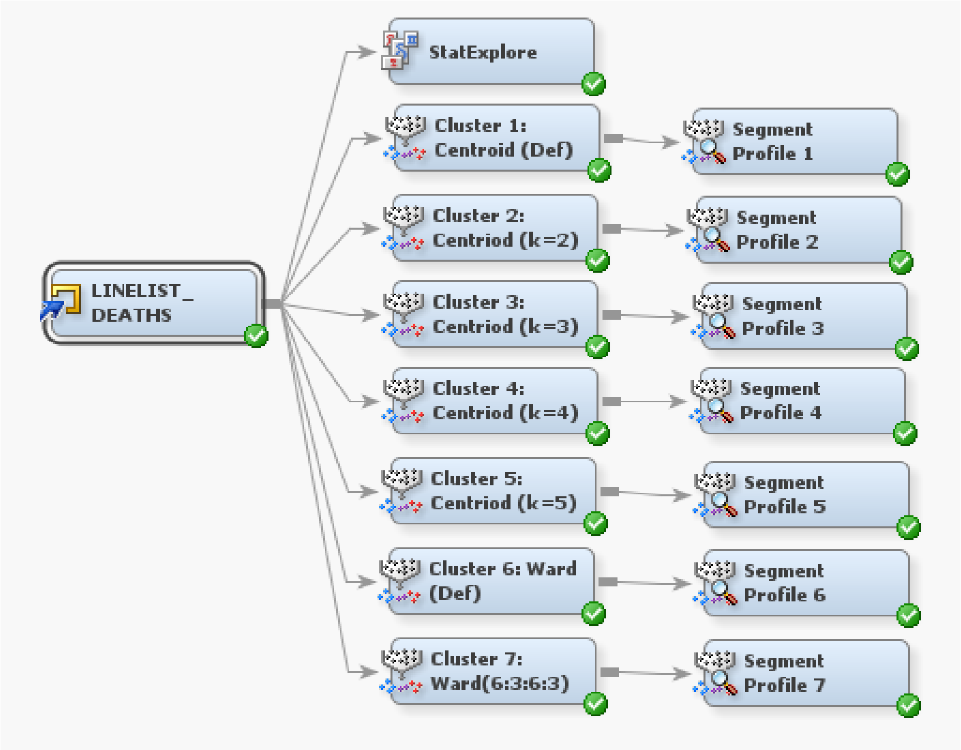


Figure 4-1. Clustering Analysis Diagram Flow in SAS Enterprise Miner

4.2 Time Series Analysis

In this study, the time series analysis was carried out using SAS Enterprise Miner. However, due to the constraint in SAS Enterprise Miner, the method used for time series analysis in this study is Exponential Smoothing. It is an approach that is similar to the common time series analysis method (Moving Average) but the difference is that exponential smoothing approach uses smoothing factor to determine how fast the weight decrease for previous observation [47]. Figure 4-2 displays the diagram flow of time series analysis in this study.

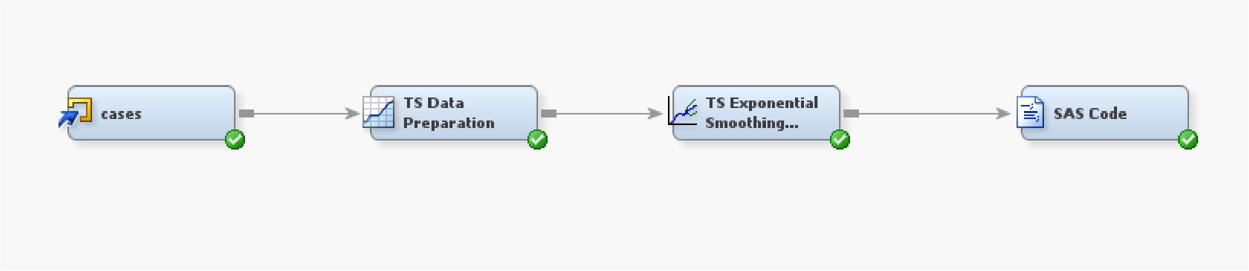


Figure 4-2. Time Series Analysis Diagram Flow in SAS Enterprise Miner

Firstly, the cases Malaysia dataset was imported to SAS Enterprise Miner using file import node. Secondly, TS Data Preparation node was linked from File Import node, along with a few properties changes. Schubert and Lee [48] mentioned that it was important to prepare and organise data correctly such as choosing the correct time intervals as time series is a sequence measured by uniformly spaced time intervals. In SAS Enterprise Miner, the time interval is automatically detected and select by default on TS Data Preparation node and TS Exponential Smoothing node [49] [50]. However, this study focuses on forecasting daily cases, thus, day was selected as the time interval for both TS Data Preparation node and TS Exponential Smoothing Node. The start and end date were manually specified from 25th January 2020 to 31st October 2021 because this analysis aims to forecast the daily cases in November and December 2021. Furthermore, the forecast led was changed to 61, meaning that the model will predict the daily cases from November 2020 until December 2021 (61 days). The properties for TS Data Preparation node (refer to Figure 4-3) and TS Exponential Node (refer to Figure 4-4) are shown in the appendix section. Finally, SAS Code node is linked from TS Exponential Smoothing node to export the predicted value of daily cases for further comparison/evaluation. The SAS code was referenced from [51], and it is shown in Figure 4-5 below in the appendix section.

4.3 Decision Tree

Similarly, Decision Tree has been carried out using SAS Enterprise Miner. Figure 4.3.1 shows the flow of how the decision tree is modelled. Initially, a dataset consisting attributes shown in Table 4.3.1 was imported to perform modelling with decision tree. Subsequently, File Import node was connected to three different Data Partition nodes of training and validation data on 70:30, 80:20, and 90:10 partitions. The diagram flow for this modelling process is shown in Figure 4-6.

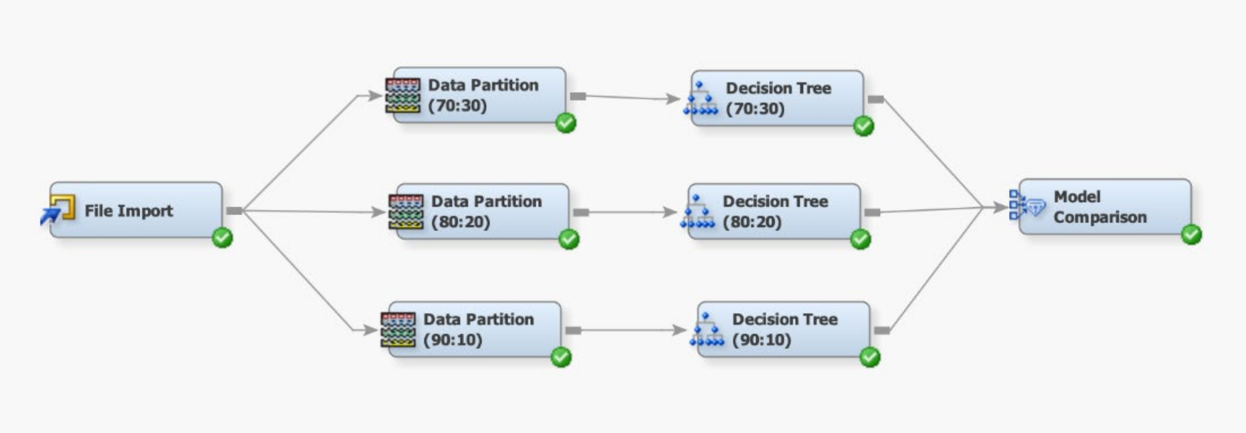


Figure 4-6. Decision Tree Diagram Flow in SAS Enterprise Miner

Table 4-3. Attributes used for Decision Tree Modelling

Attributes	**Description**
daily_case	Daily positive cases of COVID-19
daily_deaths	Daily death cases of COVID-19
daily_full	Daily number of people fully vaccinated
daily_partial	Daily number of people partially vaccinated
daily_rec	Daily recovered cases of COVID-19
date	Date
rnaught (Target)	Reproduction number of COVID-19

Consequently, the Decision Tree nodes are connected with the respective Data Partition nodes, and the properties for Decision Tree node (refer to Figure 4-7) were shown in the appendix section. As a result, with the use of Model Comparison node, the Decision Tree with 70:30 split of training and validation data produced the best model.

5. Evaluation

5.1 Objective 1: Situation of COVID-19 in Malaysia

5.1.1 Descriptive Analysis: Confirmed Cases by State

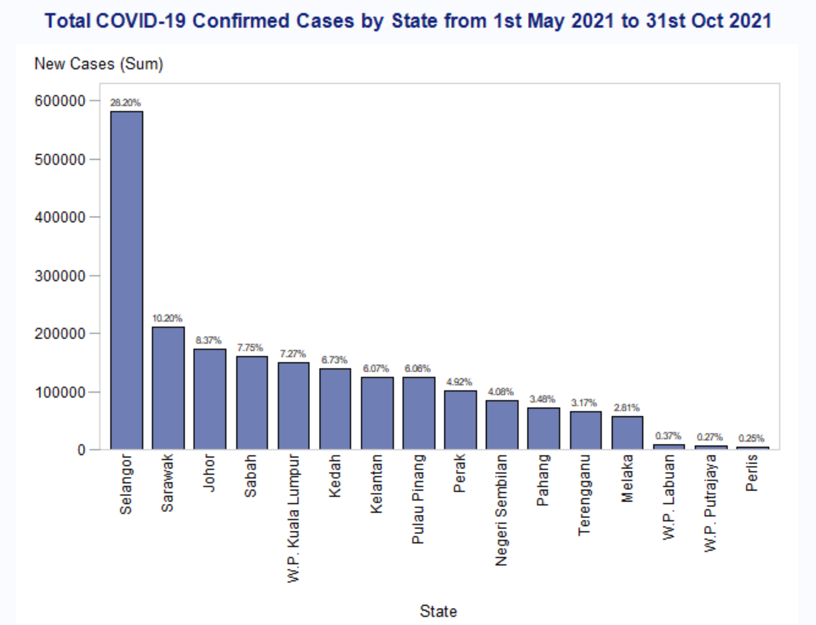
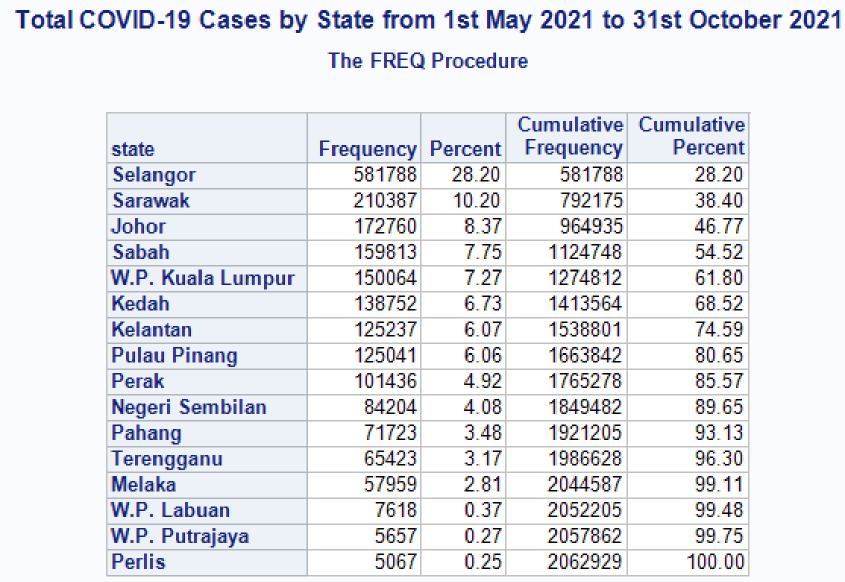


Figure 5-1. Bar Chart of Total COVID-19 Confirmed Cases by State



Total COVID-19 Cases by State from 1st May 2021 to 31st October 2021

The FREQ Procedure

state	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Selangor	581788	28.20	581788	28.20
Sarawak	210387	10.20	792175	38.40
Johor	172760	8.37	964935	46.77
Sabah	159813	7.75	1124748	54.52
W.P. Kuala Lumpur	150064	7.27	1274812	61.80
Kedah	138752	6.73	1413564	68.52
Kelantan	125237	6.07	1538801	74.59
Pulau Pinang	125041	6.06	1663842	80.65
Perak	101436	4.92	1765278	85.57
Negeri Sembilan	84204	4.08	1849482	89.65
Pahang	71723	3.48	1921205	93.13
Terengganu	65423	3.17	1986628	96.30
Melaka	57959	2.81	2044587	99.11
W.P. Labuan	7618	0.37	2052205	99.48
W.P. Putrajaya	5657	0.27	2057862	99.75
Perlis	5067	0.25	2062929	100.00

Figure 5-2. Frequency Table of Total COVID-19 Confirmed Cases by State

Malaysia has recorded a total of 2,062,929 COVID-19 cases from 1st May 2021 to 31st October 2021. From Figure 5-1, it can be observed that Selangor is the worst affected state as it reported 581,788 cases (highest) contributing to 28.20% of the overall cases in Malaysia. This is because Selangor is the densest state in Malaysia with around 6,538,000 populations. It has also fuelled the surge in daily infections in May amid the detection of new variants [52]. Sarawak and Sabah had made up 17.95% of the total cases. Perlis, WP Labuan and WP Putrajaya had cases below 10,000 which contribute to only 0.89% of the overall cases.

5.1.2 Descriptive Analysis: Death Cases by State

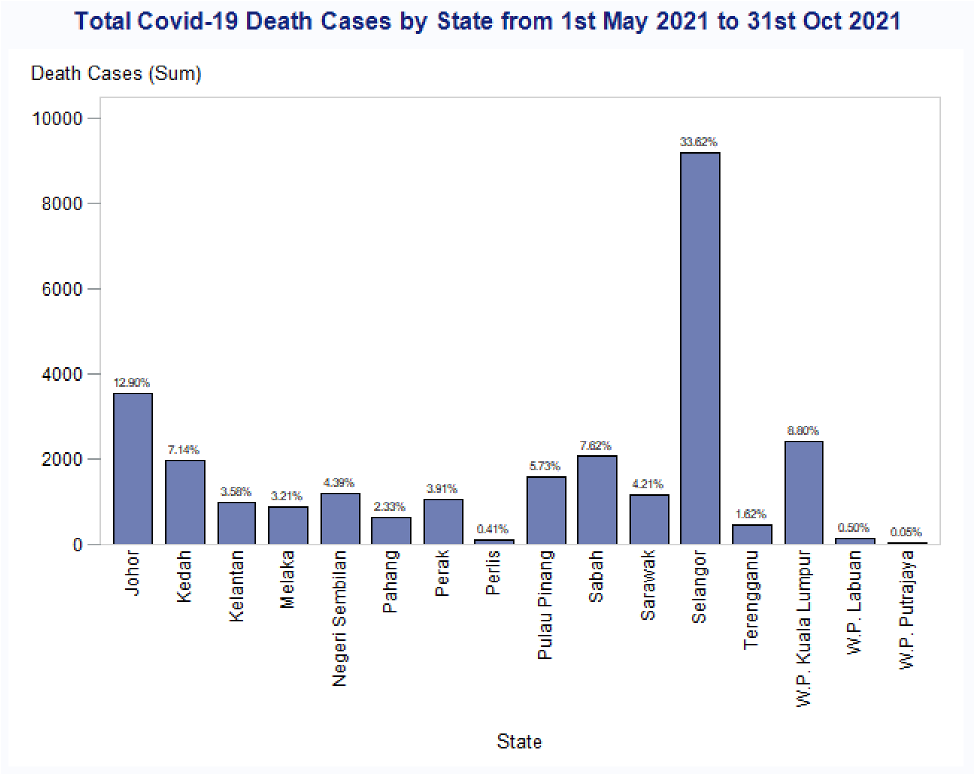
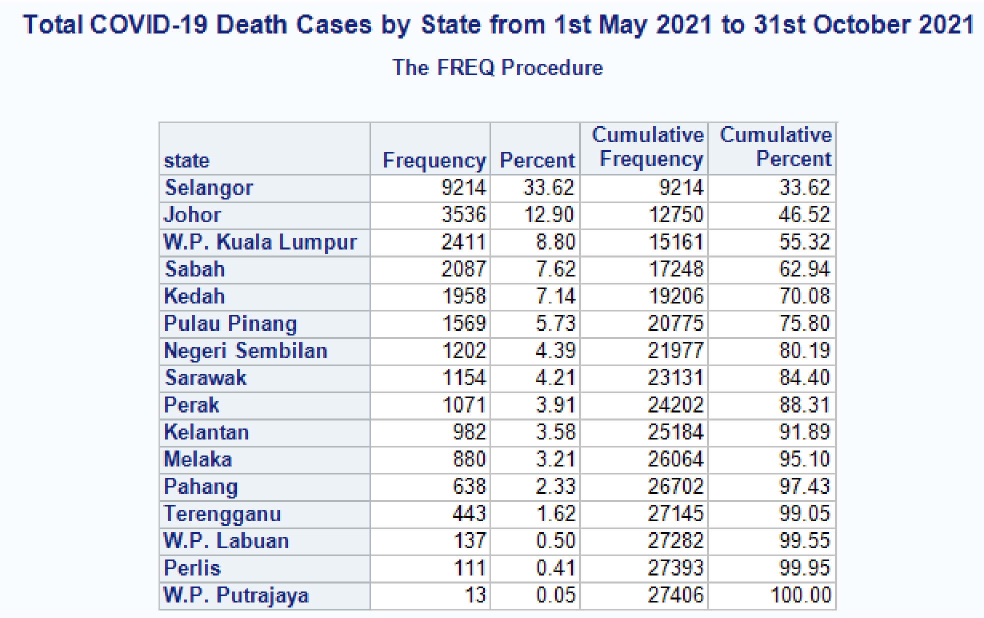


Figure 5-3. Bar Chart of Total COVID-19 Death Cases by State



Total COVID-19 Death Cases by State from 1st May 2021 to 31st October 2021

The FREQ Procedure

state	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Selangor	9214	33.62	9214	33.62
Johor	3536	12.90	12750	46.52
W.P. Kuala Lumpur	2411	8.80	15161	55.32
Sabah	2087	7.62	17248	62.94
Kedah	1958	7.14	19206	70.08
Pulau Pinang	1569	5.73	20775	75.80
Negeri Sembilan	1202	4.39	21977	80.19
Sarawak	1154	4.21	23131	84.40
Perak	1071	3.91	24202	88.31
Kelantan	982	3.58	25184	91.89
Melaka	880	3.21	26064	95.10
Pahang	638	2.33	26702	97.43
Terengganu	443	1.62	27145	99.05
W.P. Labuan	137	0.50	27282	99.55
Perlis	111	0.41	27393	99.95
W.P. Putrajaya	13	0.05	27406	100.00

Figure 5-4. Frequency Table of Total COVID-19 Death Cases by State

Malaysia has recorded a total of 27,406 COVID-19 death cases from 1st May 2021 to 31st October 2021. With Selangor reporting the most confirmed cases in Malaysia, the death cases were indeed the highest among other states. Based on Figure 5-3, death cases in Selangor were triple or more of other states with 9,214 death cases (33.62%) reported. Besides, Johor has also accounted for 12.90% of death cases, ranking the second highest in Malaysia, followed by Kuala Lumpur ranking the third which infers that the epidemic in Johor ought to be focused on.

5.1.3 Descriptive Analysis: Malaysia's COVID-19 Death Rate in 2021

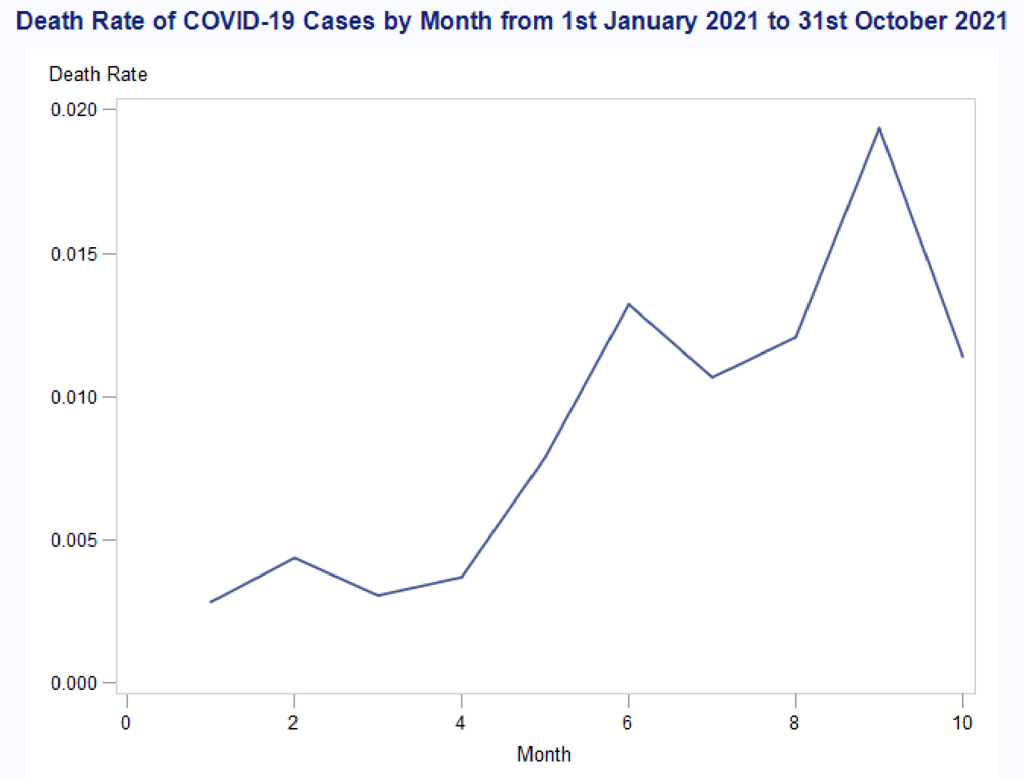


Figure 5-5. Line Chart of COVID-19 Death Rate by Month

The death rate of COVID-19 by month has spiked from January up to September with a death rate of 0.02. The first spike in April may be due to the emergence of clusters at vaccination centres along with the rapid uptake of national immunisation programs [53]. The second spike in July may be a result of the detection of a more infectious COVID-19 Delta variant [54] and state governments relaxing social distancing restrictions [55]. Nevertheless, the death rate starts to drop in October, inferring that the nations are vigilant in dealing with the COVID-19 pandemic.

5.1.4 Descriptive Analysis: Characteristics of COVID-19 Death Patients

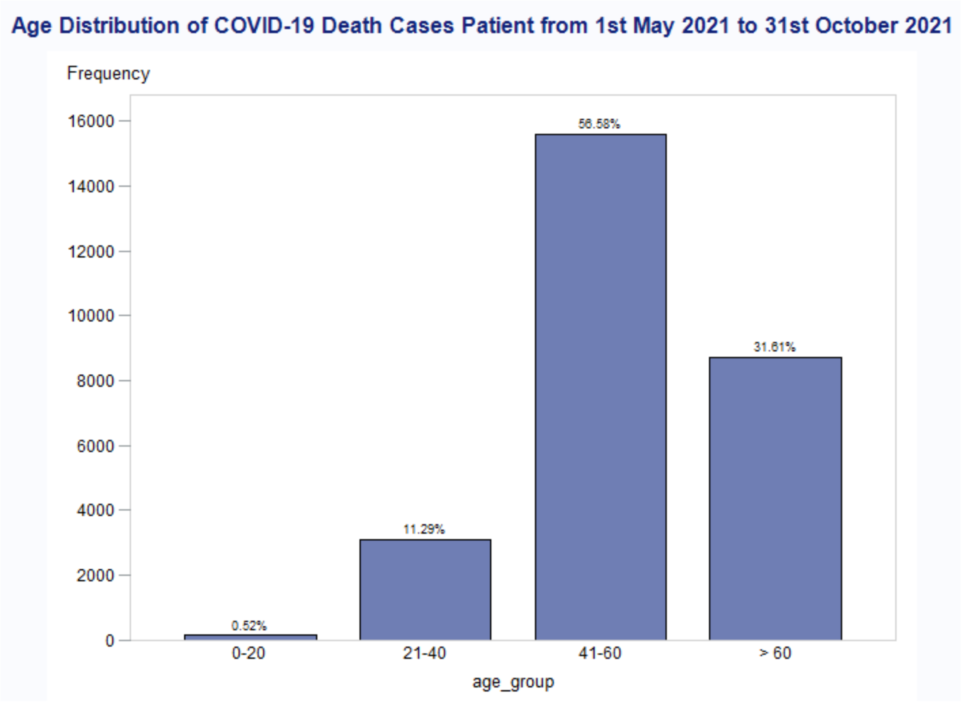


Figure 5-6. Bar Chart of Age Distribution for COVID-19 Death Patients

As observed in Figure 5-6, more than 50% of the COVID-19 deaths occurred in people over age of 40. The number of deaths in this age group is 8 times higher than that in other age groups. Research [56] showed that older adults and people with pre-existing medical conditions such as comorbidities were appeared to be more vulnerable to becoming severely ill and causing death after being infected by the COVID-19. Hence, it can be the probable reason for these two age groups having a higher death rate as compared to the others.

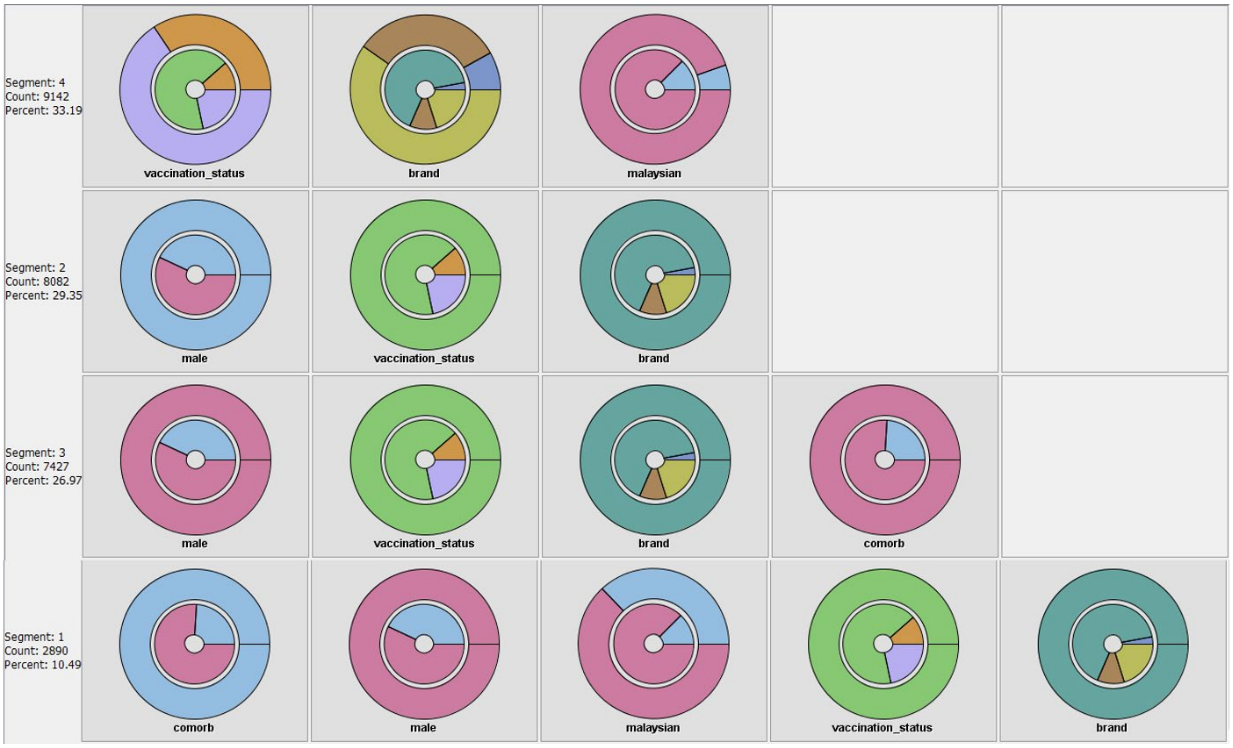


Figure 5-7. Clustering Analysis of COVID-19 Death Patients' Characteristics

The result of clustering analysis in Figure 5-7 shows that there were a total of 27,541 deaths recorded from 1st May 2021 to 31st October 2021, with 66.72% of them not vaccinated, 21.90% partially vaccinated, and 11.37% fully vaccinated. This implies that vaccine breakthrough deaths were relatively rare compared to unvaccinated people.

Segment 4 contains partially and fully vaccinated individuals which account for 9142 (33.19%) of the total death's population. Among the COVID-19 death patients, 65.80% were partially vaccinated while 34.20% were fully vaccinated. It is observed that the vaccination brand taken by the death patients varies. In comparison, Sinovac recipients account for 59.58% of deaths; Pfizer recipients account for 32.07% of deaths; AstraZeneca recipients account for 17% of death. According to [57], the greatest number of deaths among those fully vaccinated was of those inoculated with Sinovac vaccine. It is unclear whether this is due to differences in vaccine effectiveness as other factors might influence such as living conditions, occupations, socioeconomic status, education level and COVID-19 awareness.

Segment 2 contains female death patients who were not vaccinated.

Segment 3 and Segment 1 are having contrasting characteristics. Despite both segments containing all males without vaccinated, there is the presence of comorbidities in the death patients of Segment 3. Out of the 2890 male death patients in Segment 1 that do not have comorbidities, 36.81% were non-Malaysian while 63.18% were Malaysians. Common comorbidities such as hypertension, diabetes, and cardio-cerebrovascular disease were observed to be the more significant risk factors for COVID-19 patients. Moreover, preliminary data suggest that older males with comorbidities may be at higher risk for severe illness from COVID-19 [58].

5.1.5 Predictive Analysis: Time Series Model

After running the TS Exponential Smoothing node in SAS Enterprise Miner, the results generated are shown in Figure 5-8, 5-9 and 5-10.

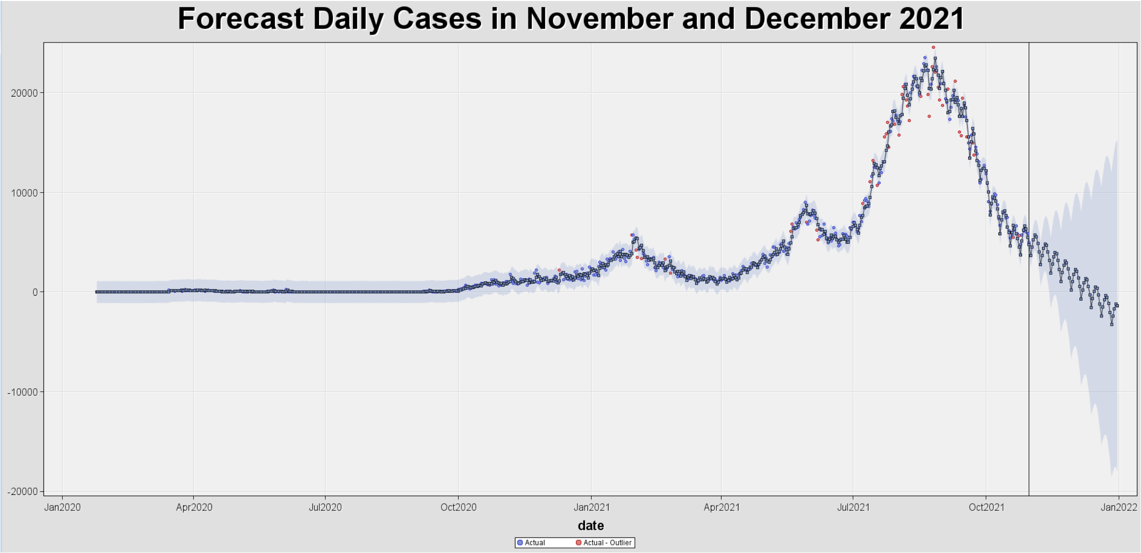


Figure 5-8. Time Series Forecast Model Result

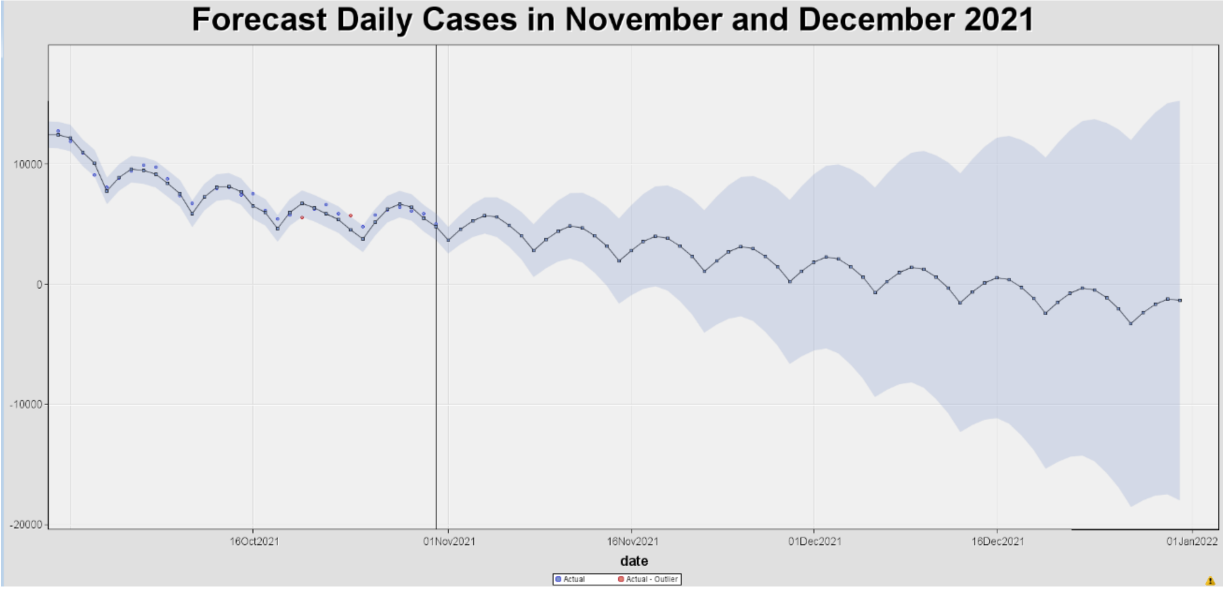
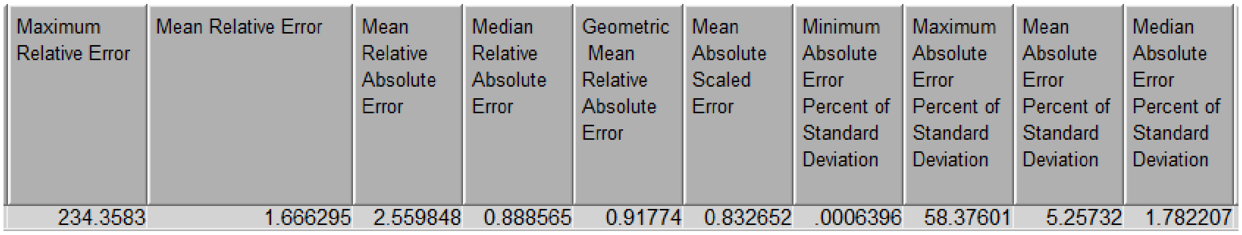


Figure 5-9. Time Series Forecast Model Result (Zoomed)



Maximum Relative Error	Mean Relative Error	Mean Relative Absolute Error	Median Relative Absolute Error	Geometric Mean Relative Absolute Error	Mean Absolute Scaled Error	Minimum Absolute Error Percent of Standard Deviation	Maximum Absolute Error Percent of Standard Deviation	Mean Absolute Error Percent of Standard Deviation	Median Absolute Error Percent of Standard Deviation
234.3583	1.666295	2.559848	0.888565	0.91774	0.832652	.0006396	58.37601	5.25732	1.782207

Figure 5-10. Time Series Forecast Model Fit Statistics

The fit statistics (refer to Figure 5-10) shows that the Mean Relative Error for the Time Series model was 1.67%. In the study by [36], the author claimed that the study's model (1.6% Mean Relative Error) is good and can perform well in predicting future value. Therefore, it can be concluded that the Time Series Model in this study can predict well in predicting the number of daily COVID-19 cases in Malaysia. To further evaluate the accuracy of the model, a comparison between the predicted value, actual daily cases in November, upper confidence level value, and lower confidence level value were tabulated in Table 5-1.

Table 5-1. Comparison of Actual, Predicted, Lower Limit, and Upper Limit of COVID-19 cases in November 2021

Date	Actual	Predicted	Lower Limit	Upper Limit
1st November 2021	4626	3645.994	2545.566	4746.421
2nd November 2021	5071	4540.092	3313.470	5766.715
3rd November 2021	5291	5275.459	3914.414	6636.504
4th November 2021	5713	5709.462	4206.450	7212.474
5th November 2021	4922	5558.386	3906.420	7210.353
6th November 2021	4701	4907.713	3100.265	6715.161
7th November 2021	4343	4022.849	2053.773	5991.924
8th November 2021	4543	2779.343	584.6570	4974.028
9th November 2021	5403	3673.441	1310.009	6036.873
10th November 2021	6243	4408.808	1870.848	6946.769
11th November 2021	6323	4842.811	2124.811	7560.811
12th November 2021	6517	4691.735	1788.421	7595.050
13th November 2021	5809	4041.062	947.364	7134.759
14th November 2021	5162	3156.197	-132.768	6445.162
15th November 2021	5143	1912.691	-1625.470	5450.852
16th November 2021	5413	2806.790	-933.244	6546.824
17th November 2021	6288	3542.157	-404.4220	7488.736
18th November 2021	6380	3976.160	-181.4890	8133.809
19th November 2021	6355	3825.084	-548.030	8198.199
20th November 2021	5859	3174.411	-1418.440	7767.265
21st November 2021	4854	2289.546	-2527.210	7106.307
22nd November 2021	4885	1046.040	-4042.080	6134.162
23rd November 2021	5594	1940.139	-3378.030	7258.313
24th November 2021	5755	2675.506	-2876.740	8227.749
25th November 2021	6144	3109.509	-2680.730	8899.747
26th November 2021	5501	2958.433	-3073.640	8990.506
27th November 2021	5097	2307.760	-3969.910	8585.430
28th November 2021	4239	1422.895	-5104.060	7949.852
29th November 2021	4087	179.389	-6639.710	6998.486
30th November 2021	4879	1073.488	-6000.650	8147.623

Note: *Negative values represent 0 cases*

According to the comparison table above (Table 5-1), the model can predict the daily cases accurately in early November. However, in the mid of November onwards, Malaysia seems to have a gradually slow downtrend of COVID-19 cases. Nevertheless, the actual number of COVID-19 cases in November was either close to the model prediction value or it was within the confidence level (lower or upper). Therefore, this comparison has demonstrated and proved that the Time Series model in this study can perform well in predicting the number of COVID-19 cases in Malaysia.

After validating the accuracy of the model using November's data, the prediction of daily COVID-19 cases was extended to December 2021 as shown in Figure 5-8 and 5-9. The cone (light blue) represents the variation in the model, expanding as time passes. In other words, the daily cases up until December is on an uptrend or downtrend. Moving on, the results show that in the end of December 2021, the number of COVID-19 cases in Malaysia is around -1374 cases (0 case since it is impossible to have negative cases). It is also possible to be -17994 (0 case) or it could be as high as 15,245 cases by the end of December 2021. The result, therefore, infers that Malaysia could be free of covid on 17th December 2021 onwards if the current trend can be continued. The upper limit of 15,245 cases anticipates the potential decrease of the nation's vigilance on the imposed SOPs to result in an unexpected uptrend.

5.2 Objective 2: Effectiveness of Vaccines

5.2.1 Descriptive Analysis: Confirmed Cases by Vaccination Status

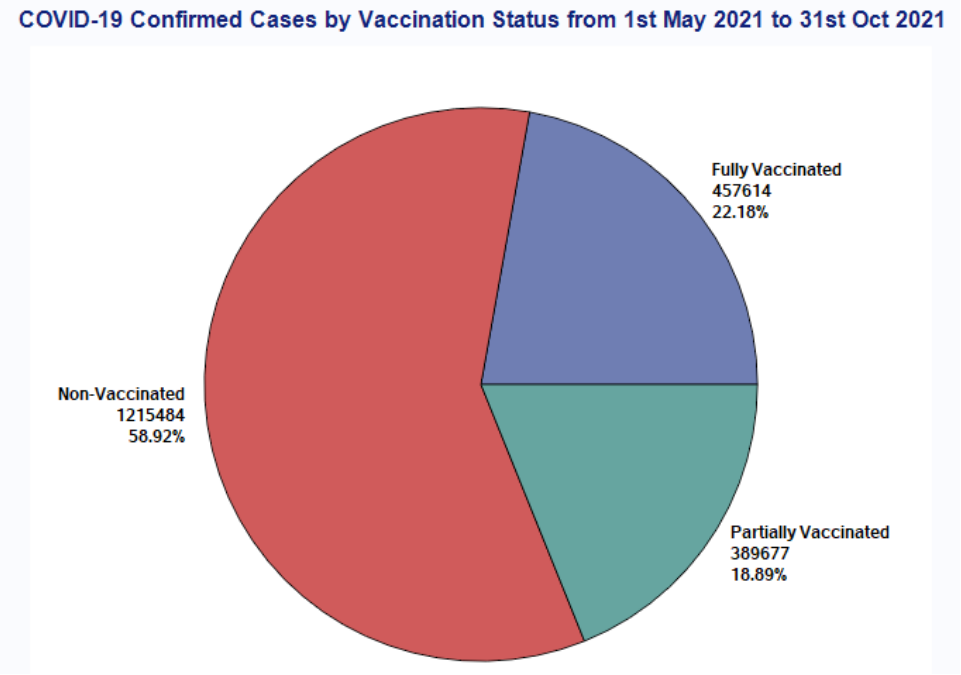


Figure 5-11. Pie Chart of Confirmed Cases by Vaccination Status

Figure 5-11 shows the total COVID-19 cases from 1st May 2021 to 31st October 2021 according to vaccination status. A total of 1,220,199 cases were reported to non-vaccinated individuals while the reported cases seem to reduce drastically in fully and partially vaccinated individuals with 450,578 and 392,152 cases respectively. The high proportion of cases among the non-vaccinated individuals is supported by [59] who confirmed that individuals who are unvaccinated and previously tested positive are 5.49 times higher chances to get re-infected compared to those vaccinated. Adding to it, unvaccinated individuals are likely to catch COVID-19 repeatedly every 16 to 17 months [60]. Hence, the mushrooming of COVID-19 cases was ascribed to the unvaccinated individuals as the infection can occur at least once among them. However, the status of vaccinated does not mean individuals are safe from the virus as there are still possible chances to get infected with COVID-19 just after vaccination. According to [61], it needs a buffer period of 14 days for the vaccine to protect the body by producing antibodies to boost the immune system. This explanation thus provided an answer to the doubt that why there are still reported cases among vaccinated individuals. Nevertheless, it is observed that the confirmed cases among fully vaccinated are higher than partially vaccinated individuals. Hence, the investigation into the death cases is another approach to confirm the effectiveness of vaccines.

5.2.2 Descriptive Analysis: Death Cases by Vaccination Status

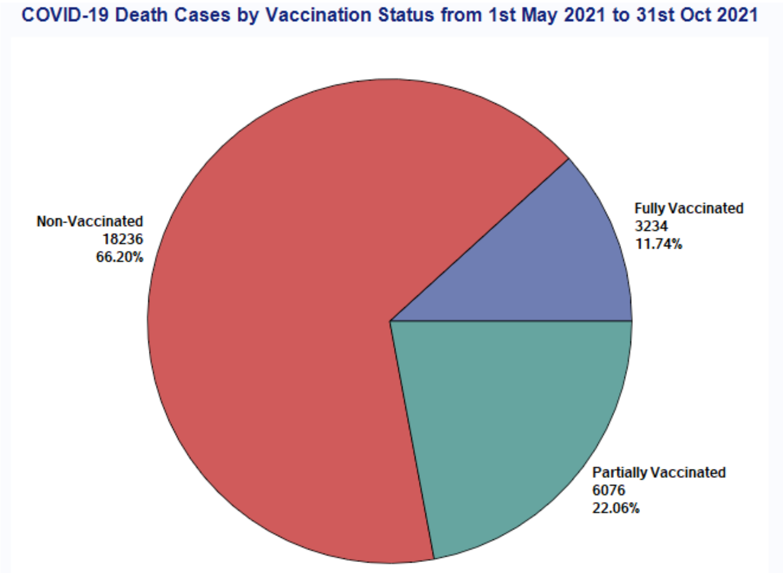


Figure 5-12. Pie Chart of Death Cases by Vaccination Status

Figure 5-12 shows the total COVID-19 death cases according to vaccination status. Death cases among non-vaccinated individuals were 3 times higher than either partially or fully vaccinated individuals which presented as the largest pie with reported 18,259 death cases (66.32%) from 1st May 2021 to 31st October 2021. According to [62], the cause of death by COVID-19 virus is due to low immune system in the body especially among the elderly. The COVID-19 virus attacks the respiratory system which can result in breathing difficulties due to inflammation in the lungs. Jandu [62] claimed that such inflammation may require a ventilator as assistance. On that account, an overloaded lung function that responds to such extreme extend to the virus can cause enormous damage to the body which is impossible for a weak immune system to sustain and thus leads to fatal consequences. To sum up the findings stated, the COVID-19 vaccine is effective in reducing mortality rate. The Public Health England (PHE) has confirmed the effectiveness of vaccines with clinical data [63]. For those in their 30s who have received both doses of vaccine, the mortality rate is reduced by 90% while the decrease is 87% and 90% for the 40s and 50s respectively.

5.2.3 Descriptive Analysis: Trend of COVID-19 Cases and Fully Vaccinated Populations

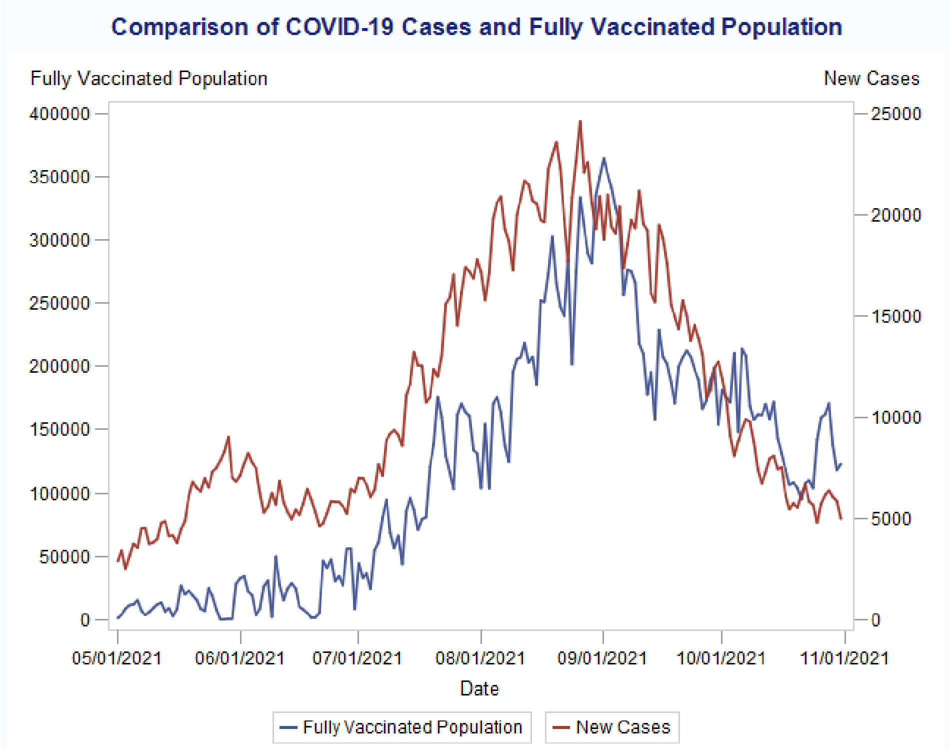


Figure 5-13. Line Chart of Comparison for Daily COVID-19 Cases and Fully Vaccinated Population

Figure 5-13 shows the comparison of new COVID-19 cases (red) and the count of fully vaccinated individuals (blue) from 1st May 2021 until 31st October 2021. The date lies on the horizontal axis has been added 14 days as the requirement for individuals to be considered fully vaccinated is 14 days after their second dose vaccine [64]. The overall performance has shown an improved situation after the increase of full vaccination rate. However, the tendency of cases after fully vaccinated and total of fully vaccinated individuals seem to be higher since mid of August 2021 and reached a peak in September 2021. According to [54], the skyrocketing of confirmed COVID-19 cases is due to the discovery highly contagious Delta variant which came right after the state of emergency was lifted in early August. Adding to it, those who were fully vaccinated or living in states under Phase 2 as well as later of NRP, have had their restrictions relaxed. Therefore, these can explain the surge of infections during these months.

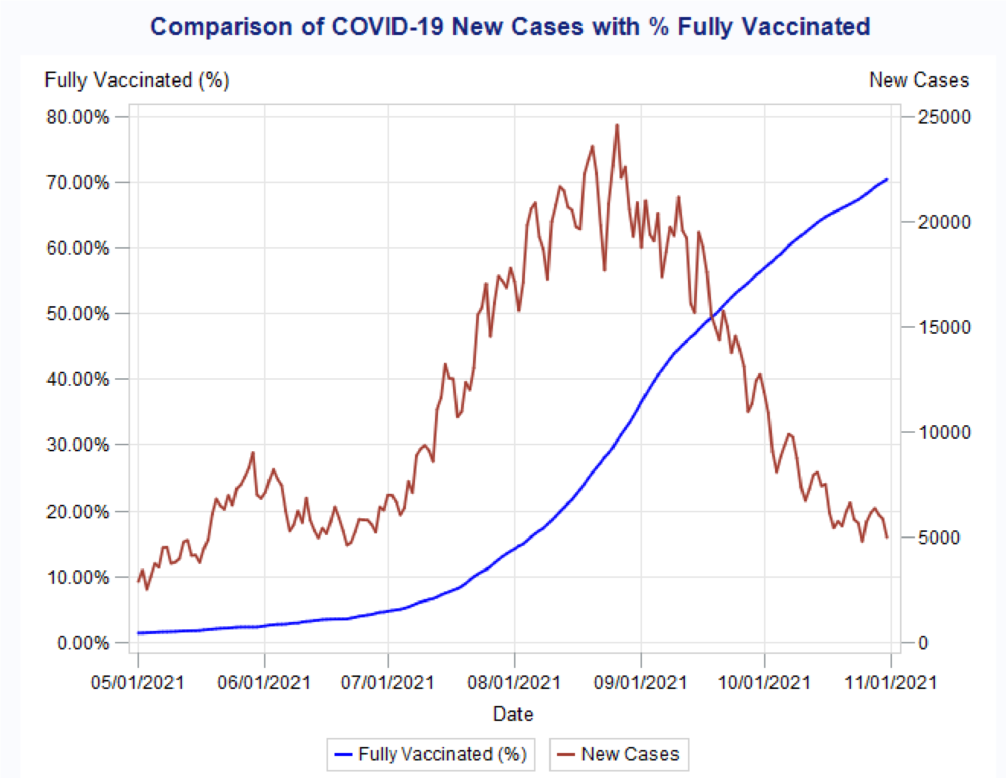


Figure 5-14. Line Chart of Comparison for Daily COVID-19 Cases and Cumulative Percentage of Fully Vaccinated Population

Figure 5-14 shows the comparison of new COVID-19 cases (red) with full vaccination percentage (blue) individuals from 1st May 2021 until 31st October 2021. Although the full vaccination rates have increased rapidly since mid-July, COVID-19 new cases were closed to 20,000 cases per day and exceeded the figure at the beginning of August. The shocking amount confirmed cases can be due to the announcement of lifted MCO in 1st of August and relaxed COVID-19 restrictions by the former prime minister which allows dining in restaurants, outdoor individual sports and interstate travel [65]. By allowing people to go out, the infection rates are then increased as people were contacting each other, thus allowing the transmission of the virus. Hence, more new cases were detected during the month.

5.2.4 Descriptive Analysis: R Naught and Fully Vaccinated Populations

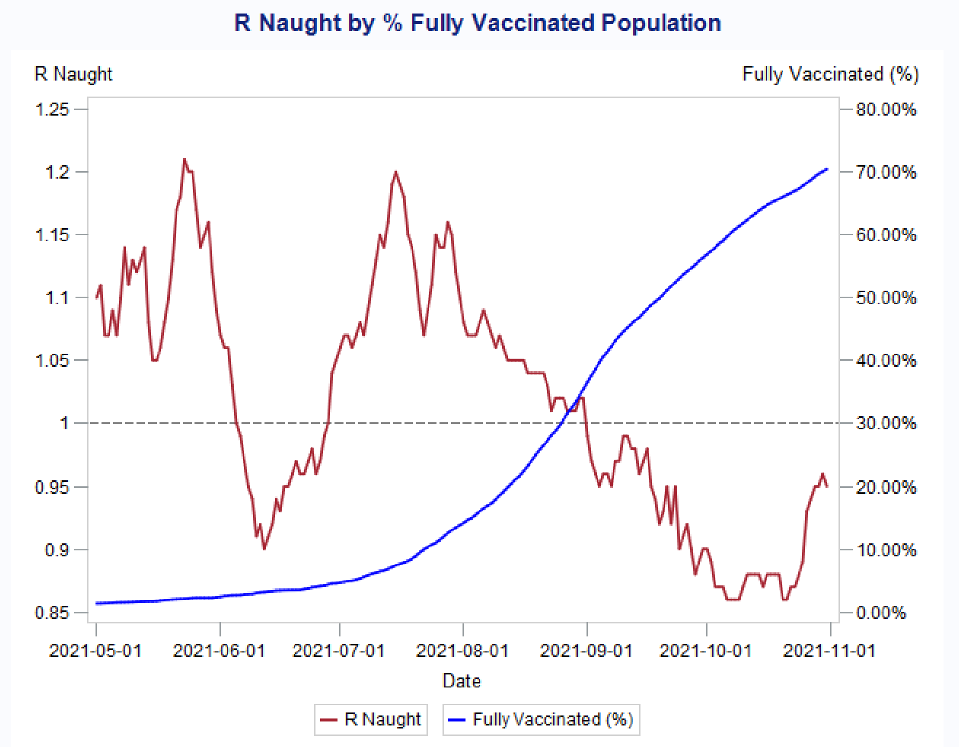
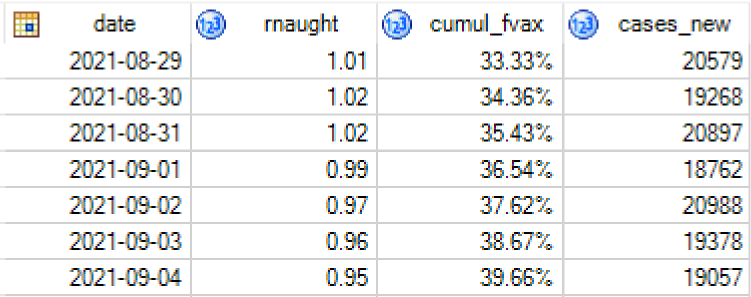


Figure 5-15. Line Chart of Comparison for Daily R Naught Value and Cumulative Percentage of Fully Vaccinated Population

Figure 5-15 shows Malaysia's R_0 value (red) along with the progression of full vaccination percentage (blue) from 1st May 2021 through 31st October 2021. The severity of COVID-19 pandemic was denoted by the fluctuation of R_0 despite more populations being fully vaccinated. Given the spike of R_0 value to reach a peak of 1.21 at the end of May 2021, the Malaysian government has commenced the Full Movement Control Order (FMCO) nationwide from 1st June 2021 onwards [66]. Such total lockdown has forced the netizens to spend most of their time at home, thus reducing the infectious rate of the disease as described by the gradual downshift of R_0 value to reach a minimum point of 0.90 during the implementation period. However, netizens tend to be less vigilant as time passes to seek social activities along with the relaxation of lockdown that allowed more economic sectors to resume their operations with FMCO progressed into phase 2 in several states [55]. Although the vaccination program was being facilitated, it was not sufficient to overcome the netizens desire to spice up their boring life. As a result, the R_0 value was forced to rise until 1.19 despite 7.10% of the population being fully vaccinated. From there onwards, the vaccination program was accelerated to a wider range of populations, resulting in an approximately 30% increase of fully vaccinated population in just 1.5 months.



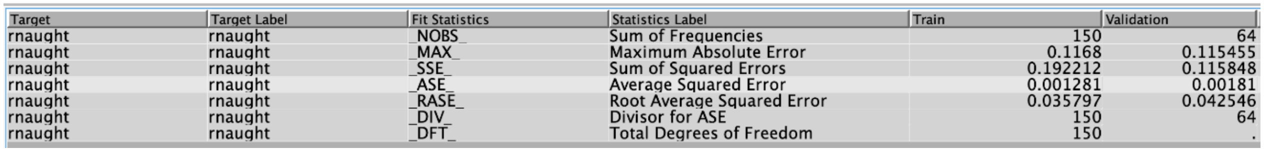
date	rnaught	cumul_fvax	cases_new
2021-08-29	1.01	33.33%	20579
2021-08-30	1.02	34.36%	19268
2021-08-31	1.02	35.43%	20897
2021-09-01	0.99	36.54%	18762
2021-09-02	0.97	37.62%	20988
2021-09-03	0.96	38.67%	19378
2021-09-04	0.95	39.66%	19057

Figure 5-16. Table of R Naught, Cumulative Fully Vaccinated Population and New Cases from 29 Aug to 4 Sep 2021

As of 1st September 2021, which was when the R_0 dropped back into the recommended safety threshold of $R_0 < 1$, 36.54% of the population has been fully vaccinated (Figure 5-16). The effectiveness of vaccine was proved with R_0 value being significantly stabilised to remain below 1 as the populations were increasingly fully vaccinated.

5.3 Objective 3: Cause for the Trend of R_0 in Malaysia

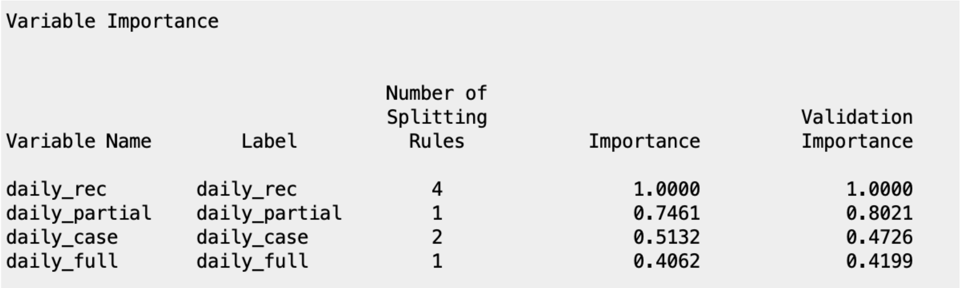
5.3.1 Predictive Analysis: Decision Tree



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
rnaught	rnaught	_NOBS_	Sum of Frequencies	150	64
rnaught	rnaught	_MAX_	Maximum Absolute Error	0.1168	0.115455
rnaught	rnaught	_SSE_	Sum of Squared Errors	0.192212	0.115848
rnaught	rnaught	_ASE_	Average Squared Error	0.001281	0.00181
rnaught	rnaught	_RASE_	Root Average Squared Error	0.035797	0.042546
rnaught	rnaught	_DIV_	Divisor for ASE	150	64
rnaught	rnaught	_DFT_	Total Degrees of Freedom	150	.

Figure 5-17. Fit Statistics of Best Decision Tree Model

As the purpose of the Decision Tree model built is to estimate the reproduction number (R_0) of COVID-19 infection, the model is then evaluated based on Average Squared Error (ASE). As a result, the model gives an ASE value of 0.002 as shown in Figure 5-17, meaning that the model can predict R_0 well with the independent variables used as mentioned in Table 4-3.



```
Variable Importance


                                 Number of
                                 Splitting                             Validation
Variable Name        Label         Rules          Importance           Importance

daily_rec        daily_rec           4               1.0000               1.0000
daily_partial    daily_partial       1               0.7461               0.8021
daily_case       daily_case          2               0.5132               0.4726
daily_full       daily_full          1               0.4062               0.4199
```

Figure 5-18. Variable or Feature Importance for Prediction

Figure 5-18 demonstrates the daily recovery cases (‘daily_rec’) as the most important factor in determining the daily R_0 value, followed by the number of people getting vaccinated partially (‘daily_partial’), daily COVID-19 positive cases (‘daily_cases’), and the number of people getting fully vaccinated (‘daily_full’). The reason being so is mostly because R_0 is usually determined by the factors such as infectious period, mode of transmission, and contact rate [30], and these factors are somehow related to the independent variables used to predict the reproduction number. For example, when a number of populations that is susceptible to be infected by COVID-19 has recovered from being COVID-19 compromised, it has impact on the contact rate of the infection, and directly cause the R_0 to decrease as expected.

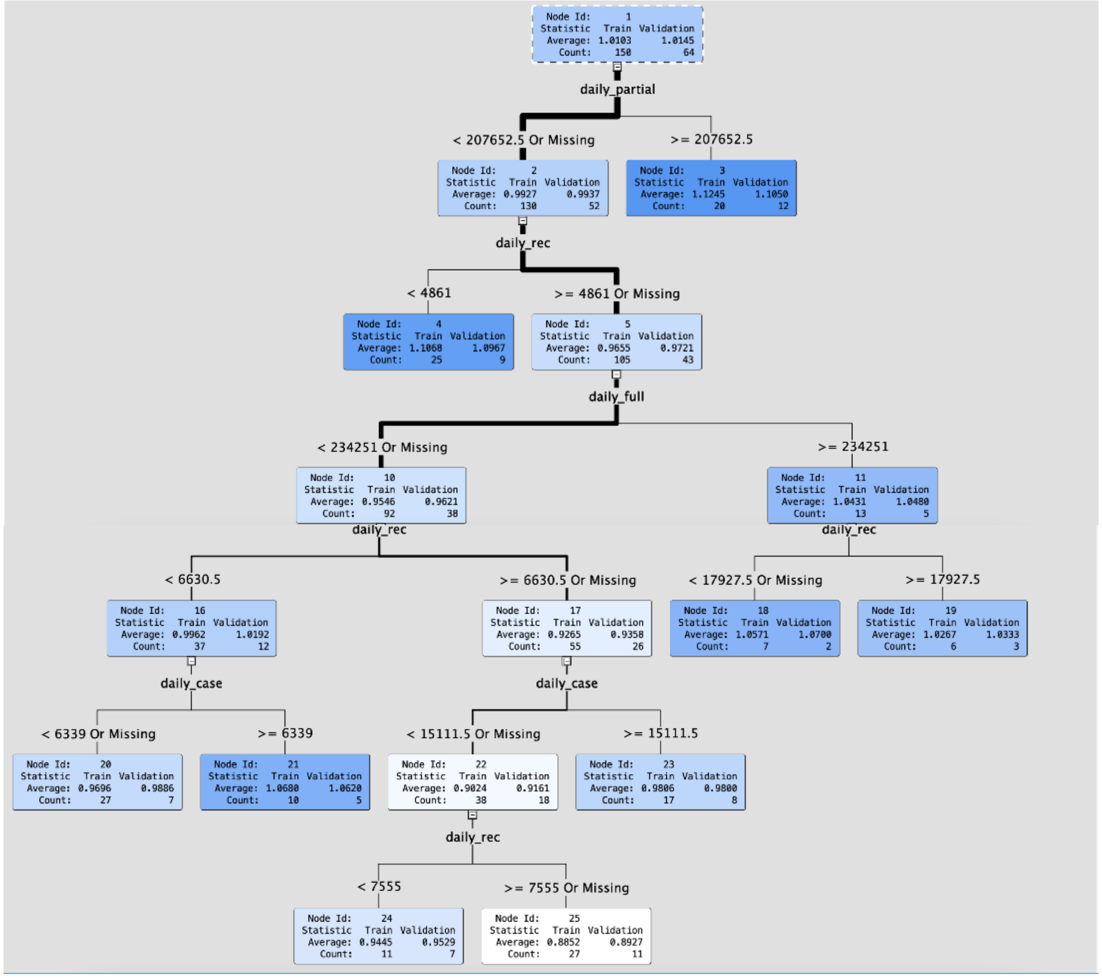
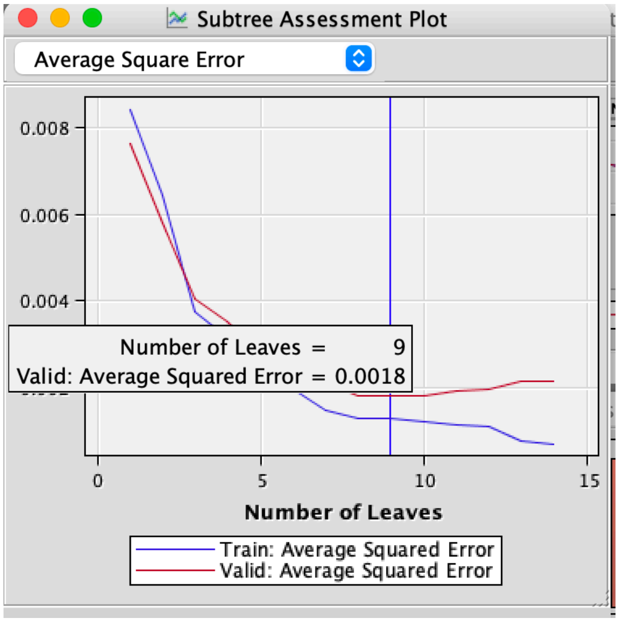

Node Id: 1 Statistic Train Validation Average: 1.0103 1.0145 Count: 150 64
daily_partial
< 207652.5 Or Missing
>= 207652.5
Node Id: 2 Statistic Train Validation Average: 0.9927 0.9937 Count: 130 52
Node Id: 3 Statistic Train Validation Average: 1.1245 1.1050 Count: 20 12
daily_rec
< 4861
>= 4861 Or Missing
Node Id: 4 Statistic Train Validation Average: 1.1068 1.0967 Count: 25 9
Node Id: 5 Statistic Train Validation Average: 0.9655 0.9721 Count: 105 43
daily_full
< 234251 Or Missing
>= 234251
Node Id: 10 Statistic Train Validation Average: 0.9546 0.9621 Count: 92 38
Node Id: 11 Statistic Train Validation Average: 1.0431 1.0480 Count: 13 5
daily_rec
daily_rec
< 6630.5
>= 6630.5 Or Missing
< 17927.5 Or Missing
>= 17927.5
Node Id: 16 Statistic Train Validation Average: 0.9962 1.0192 Count: 37 12
Node Id: 17 Statistic Train Validation Average: 0.9265 0.9358 Count: 55 26
Node Id: 18 Statistic Train Validation Average: 1.0571 1.0700 Count: 7 2
Node Id: 19 Statistic Train Validation Average: 1.0267 1.0333 Count: 6 3
daily_case
daily_case
< 6339 Or Missing
>= 6339
< 15111.5 Or Missing
>= 15111.5
Node Id: 20 Statistic Train Validation Average: 0.9696 0.9886 Count: 27 7
Node Id: 21 Statistic Train Validation Average: 1.0680 1.0620 Count: 10 5
Node Id: 22 Statistic Train Validation Average: 0.9024 0.9161 Count: 38 18
Node Id: 23 Statistic Train Validation Average: 0.9806 0.9800 Count: 17 8
daily_rec
< 7555
>= 7555 Or Missing
Node Id: 24 Statistic Train Validation Average: 0.9445 0.9529 Count: 11 7
Node Id: 25 Statistic Train Validation Average: 0.8852 0.8927 Count: 27 11

Figure 5-19. Best Decision Tree Model



Number of Leaves = 9
Valid: Average Squared Error = 0.0018

Figure 5-20. Subtree Assessment Plot of Best Decision Tree Model

Figure 5-19 shows the best tree model produced with the aforementioned independent variables. A tree model with the 9 leaves is determined as the best tree in estimating the R_0, with an ASE of 0.002, as shown in Figure 5-20. Overall, the tree model shows how different factors affect the estimation of R_0. For instance, the model suggested that if the number of people getting their first dose of vaccination or partially vaccinated in a day is less than 207652.5 and the number of recovered cases in a day is less than 4861, it is estimated that the R_0 value would be at 1.0967, on average.

6. Deployment

6.1 Strategy 1: Encourage Heterologous Booster Dose Vaccination

Based on the result in Figure 5-7, Sinovac vaccine recipients dominate the COVID-19 death cases with 20.17% as compared to Pfizer (10.84%) and AstraZeneca vaccine recipients (2.78%). It appears that the vaccine efficacy of Sinovac is low, with supporting evidence from [57] claiming that Sinovac has accounted for the greatest breakthrough death cases as compared to other types of vaccine. Therefore, Sinovac vaccine recipients should be encouraged to pursue booster dose vaccination.

However, there are only limited numbers of vaccination centres in Malaysia offering Sinovac vaccines as Pfizer and AstraZeneca vaccines are recommended as booster doses by the Technical Working Group of COVID-19 Immunisation Task Force Booster (CITF-B) [67]. Nevertheless, Sinovac vaccines are still made available in private clinics for recipients who wished, but they will have to pay for it. This has in turn discouraged people who had received their first 2 doses of Sinovac vaccines to pursue such booster dose.

With that in mind, the Malaysian government has introduced the heterologous vaccination approach that an individual will receive the vaccine of different types from the first 2 doses as their booster dose. In other words, Individuals who received first 2 doses of Sinovac vaccines have been allowed to receive Pfizer vaccines as their booster dose. Such ‘mix and match’ regimens have been proved effective in generating strong immune responses and are safe for administration [68] [69]. However, the public is generally resistant to committing to such approach due to the absence of its precedent in Malaysia.

As such, the Malaysian health authorities should publicise more information and statistics on the effectiveness of such heterologous vaccination approach to reassure the public of its safety and eliminate the fear of unknown among the netizens. They should also emphasise on encouraging Sinovac recipients to opt for Pfizer vaccines as their booster dose to jointly reduce the severity of COVID-19 pandemic in Malaysia.

6.2 Strategy 2: Allocation Healthcare Resources based on Cases Prediction

The government should ensure the adequate provision of healthcare facilities for all COVID-19 patients, but that is not the case in Malaysia, at least for now, due to the limited number of resources. Malaysian Ministry of Health (MOH) then reacted to such situation by prioritising the patients with comorbidities and are clinically unstable with the available healthcare resources, thereby transferring those who do not fall under these categories to “step down” centres for quarantine [17].

According to [70], Malaysia has been outsourcing its healthcare support services since 1996 as proposed in the Seventh Malaysia Plan. It then implies that the insufficiency of resources is caused by the misjudgement of healthcare facilities demand at various points of event, which has in turn caused the outsourcing activities to not be in the right place at the right time. Therefore, it is suggested that the authorities could review the historical trends of COVID-19 waves of severity to predict the amount of healthcare resources needed in the near future.

The results of time series analysis in this study have demonstrated good prediction of Malaysia’s COVID-19 cases in November 2021, all of which lie within the 95% confidence level of the predicted cases as shown in Table 5-1. Such accuracy indicates that time series prediction is a good measure to predict the future trend of COVID-19 cases, providing opportunities for authorities to review on a similar level of severity from the past and react accordingly.

Therefore, it is suggested that the Malaysian health authorities should make use of the time series model to predict the uptrend or downtrend of COVID-19 cases, from there derive a basis to assist them in the decision-making process of resource allocation in tackling the mounting pandemic challenges. If such measures can be deployed successfully, resources such as ICU beds and ventilators can be provisioned fairly for individuals who are in need, thus reducing the excruciation among COVID-19 patients in Malaysia.

6.3 Strategy 3: Implementation of COVID-19 Restriction based on R_0 Values

The Malaysian government has imposed various types of Movement Control Order with a common purpose to curb the spread of COVID-19 through restriction on citizens' movement and operation of the economic sector. Nevertheless, all these restrictions were soon relaxed to not compromise the country's economy. As such, it has facilitated opportunities for less vigilant individuals to seek entertainment and attend social gatherings to spice up their boring life. This has resulted in a drastic increase in COVID-19 severity in Malaysia and the government were then forced to impose further restrictions to tackle the resulting issues.

The Full Movement Control Order (FMCO), also known as total lockdown, has been enforced on 1st June 2021 for 5 weeks, then slowly progress into FMCO Phase 2 where more economic sectors are allowed to resume their businesses. Following that, Malaysia's daily new COVID-19 cases have achieved 5 digits on 13th July 2021 when restrictions were starting to be loosened on 5th July 2021 as several states progressed into FMCO Phase 2. It is supported with the result in Figure 5-15 that has demonstrated a peak of $R_0 = 1.19$ on 14th July 2021 to indicate the severity of COVID-19. The absence of further actions from the government has further promoted the spread of such virus, which has eventually recorded a new peak of 24,599 cases on 26th August 2021 [71], with the 5-digit new daily cases remained for 82 days continuously.

Therefore, the prediction of R_0 value will assist the government to respond with appropriate preventive measures inhibit any potential opportunities for the next outbreak of COVID-19 pandemic. For instance, an obvious surge in R_0 value at the end of October should suggest the government revise the Standard Operating Procedures (SOP) in November to avoid it happening as predicted. The restrictions should only be relaxed when there is a gradual downtrend of R_0 value while maintaining below the $R_0 = 1$ threshold in a stable manner. The COVID-19 pandemic in Malaysia will have a chance to come to an end if all these strategies are considered and implemented by the Malaysian government.

7. Conclusion

This study utilises the descriptive and predictive approaches in SAS Enterprise Guide and SAS Enterprise Miner to understand the situation of COVID-19, effectiveness of vaccine in reducing COVID-19 cases in Malaysia as well as the cause for the uptrend or downtrend of R_0 value. The result showed that Selangor dominates both the number of confirmed and death cases, with the death rate of the country showing a significant decrease in October 2021, and individuals who are 40 years old and above being greatly vulnerable in these circumstances. Besides, the Time Series Model shows that the confirmed cases will drop to around 0 cases by December 2021, but can also surge up to 15,245 cases. However, the results remain questionable with the limitation of the study. Moreover, despite more confirmed cases being reported among those who are fully vaccinated as compared to those who are partially vaccinated, this result has then been inverted in the perspective of death cases. Anyhow, individuals who are not vaccinated still account for more than half of the population in suffering from such disease. With an obvious downtrend of daily cases and R_0 value observed when more populations are fully vaccinated, the result from decision tree thus suggested that the trend of R_0 value is greatly affected by the number of daily recovered cases, followed by the number of partial vaccinations, confirmed cases and full vaccinations. Thus, it is suggested to encourage heterologous booster dose vaccination, allocate an appropriate amount of healthcare resources based on the predicted cases and implement COVID-19 restrictions based on predicted R_0 values.

7.1 Limitation

There are several limitations to this study. Firstly, the outbreak of COVID-19 happened at the end of 2019, which means that there is less than 2 years' data available for daily cases. After feeding in all these data into the time series model, it may still be a relatively short time series whereby the trend in the existing data may not be captured into the predicted data values. As a result, the prediction may not be truly accurate despite all the predicted values lying within the confidence interval.

Furthermore, the datasets made available do not have comprehensive attributes to support the innovation of analysis that can be conducted. The intended fields are scattered around different datasets that do not have common columns and hence cannot be merged into a single analysis.

There is also limited demographic information in the line list dataset which has impeded the happening of predictive analysis.

On top of that, this study is also limited to SAS Enterprise Guide and SAS Enterprise Miner, which has further compressed the feasibility of analysis types. These analyses, including Monte Carlo Simulation that uses predefined odds of parameters to predict future trends and SIRD model that describes a person's condition in terms of Susceptible, Infected, Recovered and Death, that can be achieved using other programming languages such as Python, cannot be conducted in such circumstance.

7.2 Justification & Potential Future Improvements

With the limitations of insufficient data mentioned above, more data should be fed in for time series analysis for the sake of higher accuracy in predictive analysis that forecasts daily cases which aid governments in taking precautions in advance to avoid an unmanageable situation. Besides, different types of time series models such as ARIMA are encouraged to explore and selection of the best model that best suits the work as this study approach has limitations with the analytics tools. Moreover, the clustering approach requires more demographic details for grouping of similar characteristics and thus leads to more accurate analysis and explanation that helps the government to detect the vast amount outbreak of confirmed cases. Lastly, death or no death cases can be indicated in daily cases in which the probability of a patient's fatality can be predicted.

8. References

[1] B. Hu, H. Guo, P. Zhou and Z. L. Shi, “Characteristics of SARS-CoV-2 and COVID-19,” *Nature Reviews Microbiology,* vol. 19, no. 3, pp. 141-154, 2021.

[2] “COVID-19 Public Health Emergency of International Concern (PHEIC) Global research and innovation forum,” World Health Organisation, 12 February 2020. [Online]. Available: https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum.

[3] “WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020,” World Health Organisation, 11 March 2020. [Online]. Available: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

[4] “Impact of COVID-19 on people's livelihoods, their health and our food systems,” World Health Organisation, 13 October 2020. [Online]. Available: https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems.

[5] K. H. D. Tang, “Movement control as an effective measure against Covid-19 spread in Malaysia: an overview,” *Journal of Public Health,* pp. 1-4, 2020.

[6] CSSEGISandData, “COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University,” Github, 2021. [Online]. Available: https://github.com/CSSEGISandData/COVID-19.

[7] “COVIDNOW in Malaysia,” COVIDNOW, 6 December 2021. [Online]. Available: https://covidnow.moh.gov.my/.

[8] “Pandemic fatigue experienced by Malaysians is cause for worry, say experts,” The Star, 20 August 2021. [Online]. Available: https://www.thestar.com.my/news/nation/2021/08/20/pandemic-fatigue-experienced-by-malaysians-is-cause-for-worry-say-experts.

[9] A. Aassve, G. Alfani, F. Gandolfi and M. L. Moglie, “Epidemics and trust: The case of the Spanish Flu,” *Health Economics,* vol. 30, no. 4, pp. 840-857, 2021.

[10] O. Jarus, “20 of the worst epidemics and pandemics in history,” Live Science, 15 November 2021. [Online]. Available: https://www.livescience.com/worst-epidemics-and-pandemics-in-history.html.

[11] D. Flecknoe, B. C. Wakefield and A. Simmons, “Plagues & wars: the ‘Spanish Flu’ pandemic as a lesson from history,” *Medicine, Conflict and Survival,* vol. 34, no. 2, pp. 61-68, 2018.

[12] “2009 H1N1 Pandemic,” Centers for Disease Control and Prevention, [Online]. Available: https://www.cdc.gov/flu/pandemic-resources/2009-h1n1-pandemic.html.

[13] K. Hickok, “How does the COVID-19 pandemic compare to the last pandemic?,” Live Science, 19 March 2020. [Online]. Available: https://www.livescience.com/covid-19-pandemic-vs-swine-flu.html.

[14] M. Chan Yeung and R. H. Xu, “SARS: epidemiology,” *Respirology,* vol. 8, no. 1, pp. 9-14, 2003.

[15] LeDuc and J. W., “SARS, the First Pandemic of the 21st Century,” *Emerging Infectious Disease,* vol. 10, no. 11, p. 26, 2004.

[16] “Severe Acute Respiratory Syndrome (SARS),” World Health Organisation, [Online]. Available: https://www.who.int/health-topics/severe-acute-respiratory-syndrome#tab=tab_1.

[17] J. H. Hashim, M. A. Adman, Z. Hashim, M. F. M. Radi and C. K. Soo, “COVID-19 Epidemic in Malaysia: Epidemic Progression, Challenges, and Response,” *Frontiers in Public Health 9,* 2021.

[18] A. U. M. Shah, S. N. A. Safri, R. Thevadas, N. K. Noordin, A. A. Rahman, Z. Sekawi, A. Ideris and M. T. H. Sultan, “COVID-19 outbreak in Malaysia: Actions taken by the Malaysian government,” *International Journal of Infectious Diseases,* vol. 97, pp. 108-116, 2020.

[19] N. Pathak, “Coronavirus and COVID-19: What You Should Know,” Web MD, 28 October 2021. [Online]. Available: https://www.webmd.com/lung/coronavirus.

[20] “Symptoms,” Centers for Disease Control and Prevention, 22 February 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.

[21] Y. Alimohamadi, M. Sepandi, M. Taghdir and H. Hosamirudsari, “Determine the most common clinical symptoms in COVID-19 patients: a systematic review and meta-analysis,” *Journal of Preventive Medicine and Hygiene,* vol. 61, no. 3, pp. 204-312, 2020.

[22] Y. C. Wu, C. S. Chen and Y. J. Chan, “The outbreak of COVID-19: An overview,” *Journal of the Chinese Medical Association,* vol. 83, no. 3, pp. 217-220, 2020.

[23] T. Singhal, “A Review of Coronavirus Disease-2019 (COVID-19),” *The Indian Journal of Pediatrics,* vol. 87, no. 4, pp. 281-286, 2020.

[24] C. M. C. Rodrigues and S. Plotkin, “Impact of Vaccines; Health, Economic and Social Perspectives,” *Frontiers in Microbiology,* vol. 11, p. 1526, 2020.

[25] CodeBlue, “Covid-19 Vaccines Cut ICU Risk By 83%, Deaths By 88% In Malaysia,” CodeBlue, 23 September 2021. [Online]. Available: https://codeblue.galencentre.org/2021/09/23/covid-19-vaccines-cut-icu-risk-by-83-deaths-by-88-in-malaysia/.

[26] K. Katella, “Comparing the COVID-19 Vaccines: How Are They Different?,” Yale Medicine, 19 November 2021. [Online]. Available: https://www.yalemedicine.org/news/covid-19-vaccine-comparison.

[27] “What are the differences between Sinovac and mRNA vaccines?,” Raffles Medical Group, [Online]. Available: https://www.rafflesmedicalgroup.com/health-resources/health-articles/what-are-the-differences-between-sinovac-and-mrna-vaccines/.

[28] N. C. Achaiah, Subbarajasetty, S. B. and R. M. Shetty, “R0 and Re of COVID-19: Can We Predict When the Pandemic Outbreak will be Contained?,” *Indian Journal of Critical Care Medicine,* vol. 24, no. 11, p. 1125, 2020.

[29] C. Anastassopoulou, L. Russo, A. Tsakris and C. Siettos, “Data-based analysis, modelling and forecasting of the COVID-19 outbreak,” *PloS one,* vol. 15, no. 3, p. e0230405, 2020.

[30] A. Gunderson and L. Woskie, “Understanding Predictions: What is R-Naught?,” Harvard Global Heath Institute, 2021. [Online]. Available: https://globalhealth.harvard.edu/understanding-predictions-what-is-r-naught/.

[31] X. S. Zhang, E. Vynnycky, A. Charlett, D. D. Angelis, Z. Chen and W. Liu, “Transmission dynamics and control measures of COVID-19 outbreak in China: a modelling study,” *Scientific Reports,* vol. 11, no. 1, pp. 1-12, 2021.

[32] S. A. Rizvi, M. Umair and M. A. Cheema, “Clustering of Countries for COVID-19 Cases based on Disease Prevalence, Health Systems and Environmental Indicators,” *Chaos Solitons Fractals,* 2021.

[33] M. Zubair, A. Iqbal, A. Shil, E. Haque, M. Hoque and I. H. Sarker, “An Efficient K-means Clustering Algorithm for Analysing COVID-19,” *International Conference on Hybrid Intelligent Systems,* pp. 422-432, 2020.

[34] B. Suharjo and N. S. Y. Utama, “K-Means Cluster Analysis of Sex, Age, and Comorbidities in the Mortalities of Covid-19 Patients of Indonesian Navy Personnel,” *Jurnal Informatika dan Sains (JISA),* vol. 4, no. 1, pp. 17-21, 2021.

[35] “Time Series Analysis,” Complete Dissertation by Statistics Solutions, [Online]. Available: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/time-series-analysis/.

[36] M. Maleki, M. R. Mahmoudi, D. Wraith and K.-H. Pho, "Time series modelling to forecast the confirmed and recovered cases of COVID-19," *Travel medicine and infectious disease,* vol. 37, p. 101742, 2020.

[37] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons & Fractals,* vol. 135, p. 109864, 2020.

[38] D. Truong, "COVID-19 New Cases Forecasting Model," Linkedin, 25 March 2020. [Online]. Available: https://www.linkedin.com/pulse/covid-19-new-cases-forecasting-model-dothang-truong-ph-d-cscp/.

[39] P. Gupta, "Decision Trees in Machine Learning," 18 May 2017. [Online]. Available: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052.

[40] "Decision Trees for Classification: A Machine Learning Algorithm," Xoriant, 7 September 2017. [Online]. Available: https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html.

[41] D. Kasilingam, S. P. S. Prabhakaran, D. K. Rajendran, V. Rajagopal, T. S. Kumar and A. Soundararaj, "Exploring the growth of COVID-19 cases using exponential modelling across 42 countries and predicting signs of early containment using machine learning," *Transboundary and Emerging Diseases,* vol. 68, no. 3, pp. 1001-1018, 2020.

[42] T. Saba, I. Abunadi, M. N. Shahzad and A. R. Khan, "Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types," *Microscopy Research and Technique,* vol. 84, pp. 1462-1474, 2021.

[43] M. Buvana and K. Muthumayh, "Prediction of COVID-19 Patient using Supervised Machine Learning Algorithm," *Sains Malaysiana,* vol. 50, no. 8, pp. 2479-2497, 2021.

[44] MoH-Malaysia, "Open data on COVID-19 in Malaysia," Github, 2021. [Online]. Available: https://github.com/MoH-Malaysia/covid19-public.

[45] Kementerian Kesihatan Malaysia, "Berita Semasa Nilai R Malaysia," Kementerian Kesihatan Malaysia, 2021. [Online]. Available: https://covid-19.moh.gov.my/kajian-dan-penyelidikan/nilai-r-malaysia.

[46] "National COVID-19 Immunisation Programme," The Special Committee for Ensuring Access to COVID-19 Vaccine Supply (JKJAV), 2021. [Online]. Available: https://www.vaksincovid.gov.my/pdf/National_COVID-19_Immunisation_Programme.pdf.

[47] "The Complete Guide to Time Series Analysis and Forecasting," Towards Data Science, 7 August 2019. [Online]. Available: https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775.

[48] S. Schubert and T. Lee, “Time Series Data Mining with SAS® Enterprise Miner™,” SAS Institute Inc, 2011. [Online]. Available: https://support.sas.com/resources/papers/proceedings11/160-2011.pdf.

[49] “Time Series Data Preparation Node,” SAS® Help Center, [Online]. Available: https://documentation.sas.com/doc/en/emref/15.1/n0rzjozmp6um16n164o5w43nsaax.htm.

[50] “Time Series Exponential Smoothing Node,” SAS® Help Center, [Online]. Available: https://documentation.sas.com/doc/en/emref/15.1/p1tz958p4xofxmn1u8b970af2x9f.htm.

[51] S. R. K. Choudhury and A. Komarraju, “Understanding Crime Pattern in United States by Time Series Analysis using SAS Tools,” SESUG, 2018. [Online]. Available: https://lexjansen.com/sesug/2018/SESUG2018_Paper-264_Final_PDF.pdf.

[52] “Malaysia Reports Record Daily Covid Cases as Pressure Mounts,” Bloomberg, 19 May 2021. [Online]. Available: https://www.bloomberg.com/news/articles/2021-05-19/malaysia-adds-a-record-6-075-new-coronavirus-cases-on-wednesday.

[53] H. Hassan, “204 workers at Malaysian vaccination centre test positive for Covid-19,” The Straits Times, 14 July 2021. [Online]. Available: https://www.straitstimes.com/asia/se-asia/vaccination-centre-in-malaysia-shut-after-staffer-tests-positive.

[54] S. Salim, “Covid-19: Malaysia reports over 630,000 cases and 7,640 deaths in August, even as 13.6 mil vaccine doses doled out,” The Edge Markets, 2 September 2021. [Online]. Available: https://www.theedgemarkets.com/article/covid19-malaysia-reports-over-630000-cases-and-7640-deaths-august-even-136-mil-vaccine-doses.

[55] A. Wong, “5 States In Malaysia Upgraded To Phase 2 Of FMCO 3.0!,” Techarp, 3 July 2021. [Online]. Available: https://www.techarp.com/business/5-states-move-phase-2-fmco/.

[56] A. Sanyaolu, C. Okorie, A. Marinkovic, R. Patidar, K. Younis, P. Desai, Z. Hosein, I. Padda, I. Padda and M. Altaf, “Comorbidity and its Impact on Patients with COVID-19,” *SN comprehensive clinical medicine,* pp. 1-8, 2020.

[57] J. L. Koh, “Covid-19 deaths among vaccinated rare, mostly Sinovac recipients,” Yahoo News, 10 September 2021. [Online]. Available: https://malaysia.news.yahoo.com/covid-19-deaths-among-vaccinated-060400766.html?guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAABFT1vJP97GHWFHVklGFt54pLoJjO60RpP3pPWyS8A7vEehgGuGOqxe_p6tlcoAH5cRZdmzXX6LMTin7YbHvleWxJtQOb4kf4LEOHyyIej4.

[58] H. Ejaz, A. Alsrhani, A. Zafar, H. Javed, K. Junaid, A. E. Abdalla, K. O. Abosalif, Z. Ahmed and S. Younas, “COVID-19 and comorbidities: Deleterious impact on infected patients,” *Journal of infection and public health,* vol. 13, no. 12, pp. 1833-1839, 2020.

[59] C. Johnson, “CDC: Unvaccinated 5 times more likely to get COVID again vs. those who are vaccinated,” News Channel 5, 1 November 2021. [Online]. Available: https://www.newschannel5.com/news/cdc-unvaccinated-5-times-more-likely-to-get-covid-again-vs-those-who-are-vaccinated.

[60] C. Crist, “Unvaccinated People Likely to Catch COVID Repeatedly,” WebMD, 25 October 2021. [Online]. Available: https://www.webmd.com/vaccines/covid-19-vaccine/news/20211024/unvaccinated-people-likely-to-catch-covid-repeatedly.

[61] “How Vaccines Work,” Centers of Disease Control and Prevention, [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/how-they-work.html.

[62] O. Bowden, “How does COVID-19 cause death? Here’s what happens in the lungs,” Global News, 12 April 2020. [Online]. Available: https://globalnews.ca/news/6805639/coronavirus-covid-19-death/.

[63] B. Walker, “How Covid-19 vaccines dramatically reduce death rates,” The News Statesman, 10 September 2021. [Online]. Available: https://www.newstatesman.com/chart-of-the-day/2021/09/how-covid-19-vaccines-dramatically-reduce-death-rates.

[64] “When You’ve Been Fully Vaccinated,” Centers for Disease Control and Prevention, 15 October 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/vaccines/fully-vaccinated.html.

[65] “Malaysia to ease COVID curbs for fully vaccinated in eight states,” Reuters, 8 August 2021. [Online]. Available: https://www.reuters.com/world/asia-pacific/malaysia-ease-covid-curbs-fully-vaccinated-eight-states-2021-08-08/.

[66] L. Loh, “Covid-19: Malaysia Goes Into Full Lockdown (FMCO) From June 1, 2021,” Tatler, 30 May 2021. [Online]. Available: https://www.tatlerasia.com/style/wellness/covid-19-malaysia-full-lockdown-total-lockdown-fmco-sops.

[67] T. A. Yusof, “Only limited PPVs offering Sinovac as booster shots,” New Straits Times, 30 November 2021. [Online]. Available: https://www.nst.com.my/news/nation/2021/11/750138/only-limited-ppvs-offering-sinovac-booster-shots.

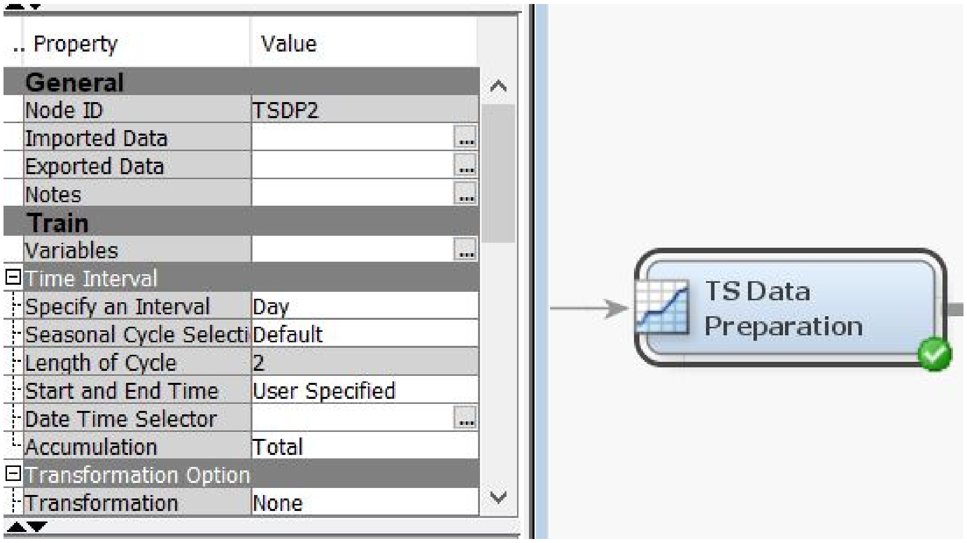
[68] V. K. D. S. J. M. F. H. S. M. A. A.-O. L. Z. C. G. Tina Schmidt, R. Urschel, S. Schneitler, S. L. Becker, B. C. Gärtner, U. Sester and M. Sester, “Immunogenicity and reactogenicity of heterologous ChAdOx1 nCoV-19/mRNA vaccination,” *Nature Medicine,* vol. 27, no. 9, pp. 1530-1535, 2021.

[69] R. H. Shaw, A. Stuart, M. Greenland, X. Liu, Van-Tam, J. S. Nguyen and M. D. Snape, "Heterologous prime-boost COVID-19 vaccination: initial reactogenicity data," *The Lancet,* vol. 397, no. 10289, pp. 2043-2046, 2021.

[70] F. D. Mustapa, M. Mustapa, F. Ismail and K. N. Ali, "Outsourcing in Malaysian healthcare support services: A study on the causes of increased operational costs," 2004.

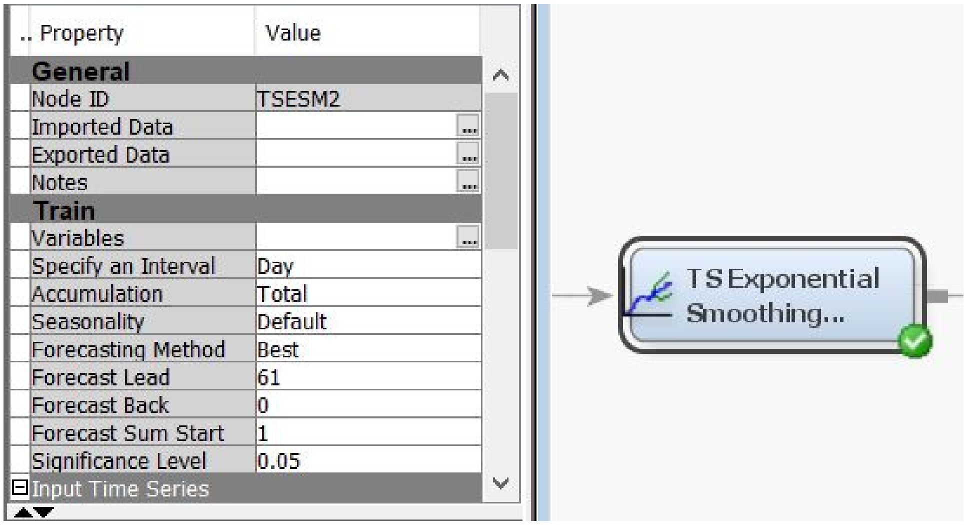
[71] J. Kaos, "Covid-19: Record high of 24,599 new cases recorded on Thursday (Aug 26)," The Star, 26 August 2021. [Online]. Available: https://www.thestar.com.my/news/nation/2021/08/26/covid-19-another-record-high-with-24599-new-cases-on-thursday-aug-26.

9. Appendix



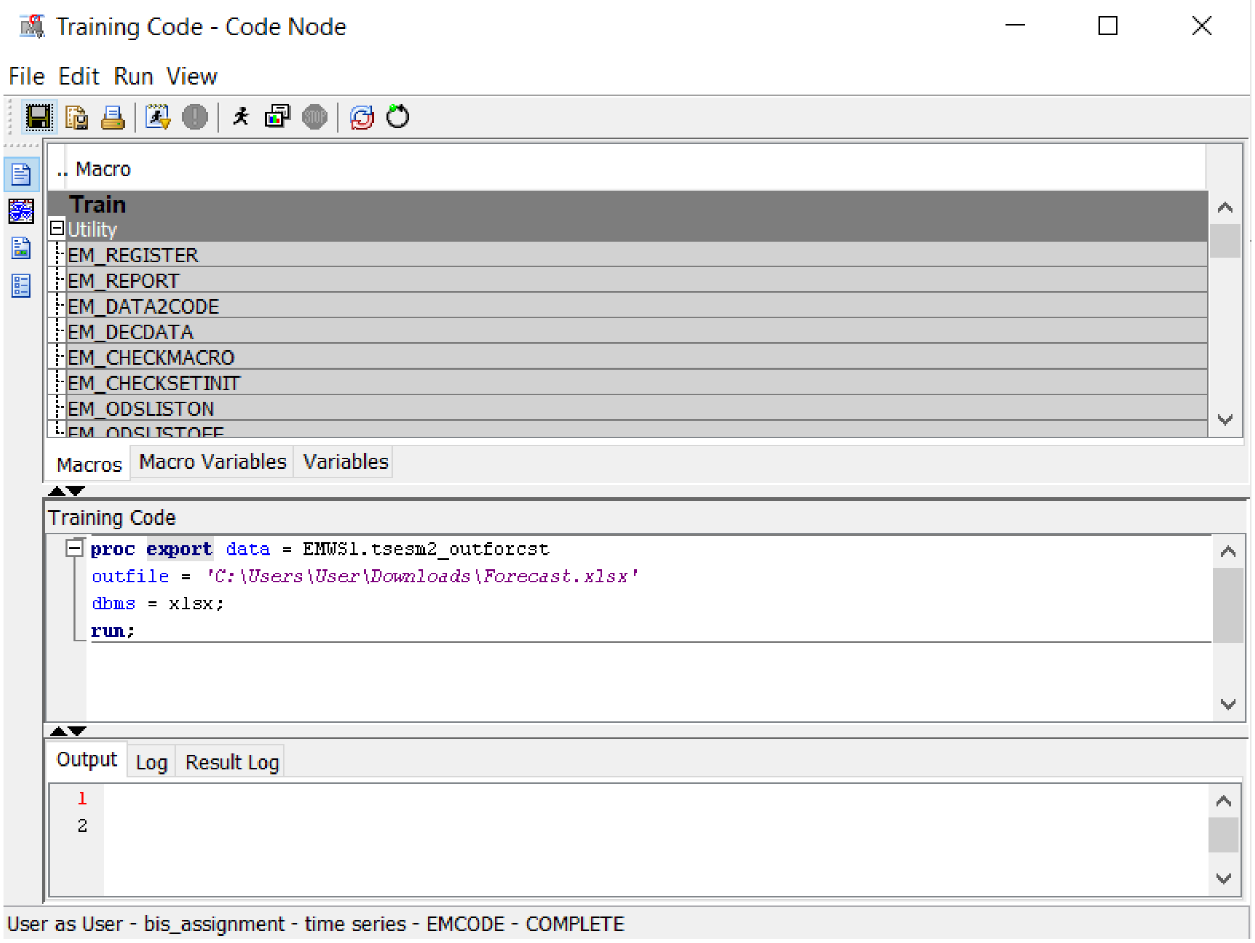
.. Property	Value
General	
Node ID	TSDP2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟Time Interval	
Specify an Interval	Day
Seasonal Cycle Selecti	Default
Length of Cycle	2
Start and End Time	User Specified
Date Time Selector	...
Accumulation	Total
⊟Transformation Option	
Transformation	None

Figure 4-3. Properties of TS Data Preparation Node



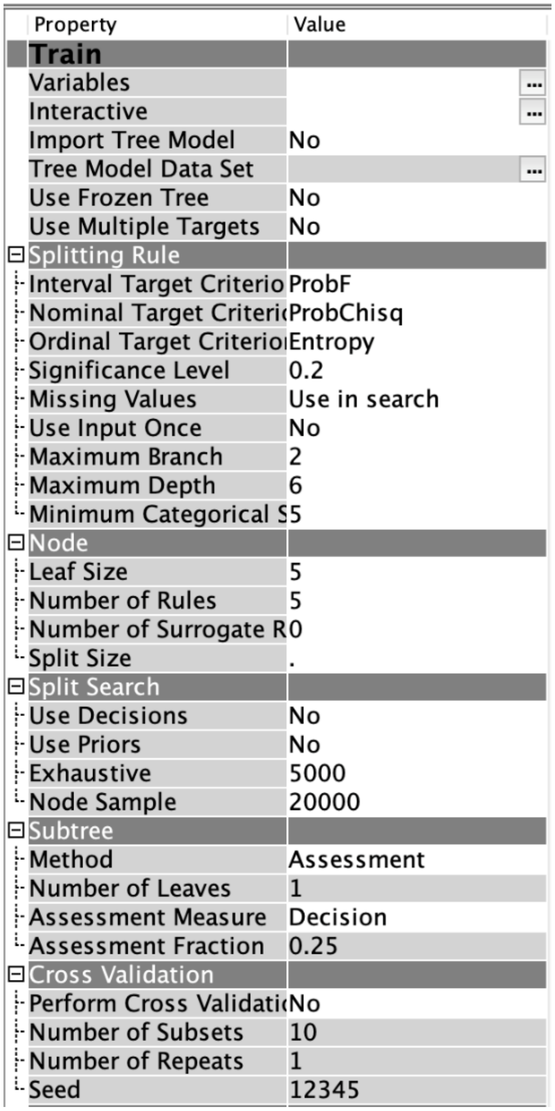
.. Property	Value
General	
Node ID	TSESM2
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Specify an Interval	Day
Accumulation	Total
Seasonality	Default
Forecasting Method	Best
Forecast Lead	61
Forecast Back	0
Forecast Sum Start	1
Significance Level	0.05
⊟Input Time Series	

Figure 4-4. Properties of TS Exponential Smoothing



```
proc export data = EMWS1.tsesm2_outforcst
outfile = 'C:\Users\User\Downloads\Forecast.xlsx'
dbms = xlsx;
run;
```

Figure 4-5. SAS Code on Exporting Time Series Forecast Result



Property	Value
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
⊟ Splitting Rule	
Interval Target Criterio	ProbF
Nominal Target Criteri	ProbChisq
Ordinal Target Criterio	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical S	5
⊟ Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate R	0
Split Size	.
⊟ Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
⊟ Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
⊟ Cross Validation	
Perform Cross Validati	No
Number of Subsets	10
Number of Repeats	1
Seed	12345

Figure 4-7. Properties of Decision Tree Nodes