UNIVERSITI PUTRA MALAYSIA

PREDICTION OF PROBABILITY OF DEATH OF COVID-19: MALAYSIAN CASE STUDY

ARINA BINTI MOHAMED

PREDICTION OF PROBABILITY OF DEATH OF COVID-19:

MALAYSIAN CASE STUDY

by

ARINA BINTI MOHAMED

Thesis Submitted to the Department of Physics, Universiti Putra Malaysia, in partial Fulfilment of the Requirements for the Degree of Bachelors of Science in Physics with Education (Honours)

February 2022

ABSTRACT

PREDICTION OF PROBABILITY OF DEATH OF COVID-19 DISEASE: MALAYSIAN CASE STUDY

by

Arina binti Mohamed

198783

February 2022

Supervisor: Dr. Mohd Amiruddin Abd Rahman

Faculty: Faculty of Science

The world was shocked when the first series of COVID-19 cases were reported in late December 2019 in Wuhan, China. Clusters of infections has spread outside China which later urged the World Health Organization (WHO) to call it a pandemic. The pandemic has not only burdened the healthcare sectors, but it has also caused the economy to crippled causing millions of people struggling to carry on with their life. When an outbreak of new and unknown disease happens, it is very important to study the factors of the infections. Also, it is very crucial to investigate the trend and predict the upcoming situation so that the government can take a better plan to lessen the risk of the infections. This project aims to compare the result based on different prediction input parameters, which are Malaysian state, patients' age and medical history. Also, this project intent to investigate the relationship between the number of death with number of cases from yesterday's up to 10 days. This lagging of days is called as lags. To achieve the objectives, a set

of new confirmed cases data and death cases data for each state, patient's age and medical history are collected and analyzed before being divided into two parts for training and testing. This project used machine learning model, ARIMAX to predict the probability of death for each of the state, patient's age and medical history. From the findings, the correlation across the data set discovered that the data set of new confirmed cases correlates with the data of death cases for each state, patient's age and medical history with average of 0.6875. The accuracy and errors for each of the death cases data for state, patients' age and medical history differ depends on the size of the training and testing data. For prediction of death in each state, W.P Labuan has recorded the lowest error, Root Mean Square Error (RMSE) and Mean Average Error (MAE) of 1.8696 and 1.4606 respectively. While the prediction of death for each of the patient's age, the class age of 25-29 has recorded the lowest value of RMSE which is 1.0718. Lastly, Dyslipidemia has recorded the lowest RMSE and MAE, 3.0629 and 2.207 respectively compared to other five diseases of the patients' medical history. Overall, the prediction of death for each disease history recorded the least average error compared to the data of each state and patient's age. Also, the trend of lags in 10 days differ depends on the size of training and testing data.

ABSTRAK

RAMALAN KEBARANGKALIAN KEMATIAN DISEBABKAN COVID-19:

KAJIAN KES MALAYSIA

Oleh

Arina binti Mohamed

198783

Februari 2022

Penyelia: Dr. Mohd Amiruddin Abd Rahman

Fakulti: Fakulti Sains

Dunia dikejutkan apabila siri pertama kes COVID-19 dilaporkan pada hujung Disember 2019 di Wuhan, China. Kes jangkitan yang telah merebak ke luar negara China telah menggesa Pertubuhan Kesihatan Dunia (WHO) untuk mengesahkannya sebagai pandemik. Pandemik bukan sahaja membebankan sektor penjagaan kesihatan, malah ia juga menyebabkan ekonomi lumpuh dan mengakibatkan berjuta-juta orang bergelut untuk meneruskan kelangsungan hidup mereka. Apabila wabak penyakit baru dan tidak diketahui berlaku, faktor yang menyebabkan jangkitan adalah amat penting untuk dikaji. Selain itu, trend dan ramalan situasi yang akan datang perlu disiasat agar kerajaan dapat merancang pelbagai cara untuk mengurangkan risiko jangkitan. Projek ini dijalankan bertujuan untuk mebandingkan keputusan input parameter ramalan yang berbeza untuk setiap negeri, umur dan sejarah penyakit pesakit. Selain itu, projek ini juga berhasrat untuk menyiasat hubungan antara bilangan kes kematian dalam satu hari dengan bilangan kes berjangkit dari hari sebelumnya sehingga 10 hari. Pertundaan hari ini

dipanggil sebagai ‘Lags”. Bagi mencapai objektif tersebut, satu set data kes jangkitan baharu dan kes kematian bagi setiap negeri, umur dan sejarah penyakit pesakit dikumpul dan dianalisis sebelum dibahagikan kepada dua bahagian untuk data latihan dan data ujian. Projek ini menggunakan pembelajaran mesin model ARIMAX untuk meramal kebarangkalian kematian bagi setiap negeri, umur pesakit dan sejarah penyakit pesakit di Malaysia. Hasil dapatan kajian, korelasi merentas set data mendapati bahawa set data kes jangkitan baharu berkolerasi dengan kes kematian bagi setiap negeri, umur dan sejarah penyakit pesakit dengan purata sebanyak 0.6875. Ketepatan dan ralat bagi setiap data kes kematian untuk negeri, umur dan sejarah penyakit pesakit adalah berbeza bergantung kepada saiz data latihan dan ujian. Bagi ramalan kematian di setiap negeri, W.P. Labuan telah merekodkan ralat terendah, RMSE dan MAE masing-masing 1.8696 dan 1.4606. Manakala ramalan kematian bagi setiap umur pesakit, kelas umur 25-29 telah mencatatkan nilai RMSE terendah iaitu 1.0718. Akhir sekali, Dislipidemia telah merekodkan RMSE dan MAE terendah, iaitu 3.0629 dan 2.207 berbanding lima penyakit lain dalam sejarah penyakit pesakit. Secara keseluruhannya, ramalan kematian bagi setiap negeri memberikan ralat purata paling rendah berbanding data umur pesakit dan sejarah perubatan. Selain itu, trend pertundaan hari dalam 10 hari berbeza bergantung pada kepada saiz data latihan dan ujian.

ACKNOWLEDGEMENT

First, I would like to express my thankfulness to Allah All-Mighty for gracing me the chance and strength to complete my final year project entitled Prediction of Probability of Death of COVID-19: Malaysian Case Study. I would like to give the most appreciation to my supervisor, Dr. Mohd Amiruddin bin Abd Rahman for all the knowledge and guidance throughout my project. Also, to Ms. Caceja Elyca Anak Bundak whom has spent her time and willingness to share knowledge and help to finish this work. This dissertation could not be finish without their guidance and help.

Next, I want to express my special and deepest appreciation to my parents (Abah & Ibu) Mohamed bin Sulaiman and Zainab binti Ab Majid for their moral support throughout my journey completing this project. Not to also forget my beloved siblings for their non-stop support and love.

To all my friends especially Ajmain, Anith, Sarah, Ain, Adrina and Syarah thank you for your never-ending support and for always reach me out throughout this journey. Not to forget my final year partners, Athirah and Amir we did manage to complete this final year project together.

Lastly, thank you to all my classmates and all the lecturers from Physics Department of Faculty of Science that help me in completing my degree. No words can describe how thankful I am for all the support and love.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

The novel coronavirus (COVID-19) was named as the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2, 2019-nCoV) due to its high homology (~80%) to SARS-CoV, which caused acute respiratory distress syndrome (ARDS) and high mortality during 2002– 2003 (Ksiazek T.G.et al, 2003). Coronaviruses are large single stranded RNA viruses that can infect humans, but also wide range of animal. It was first described by Tyrell and Bynoe in 1966 when they cultivated the viruses from patients with common colds.

In December 2019, the world was shocked by the news that there are a series of acute atypical respiratory disease occurred in Wuhan, China. It was reported that the epidemic of COVID-19 disease was initially believed to start via a zoonotic transmission associated with the seafood market in Wuhan, Hubei Province in China. Later it was identified that human-to-human transmission played a major factor in the subsequent outbreak (Mackenzie & Smith, 2020). After cluster of the pneumonia cases reported, The World Health Organization (WHO) published the first Disease Outbreak News on the virus on 5 January 2020. WHO issued a package of technical guidance to help countries detect and manage potential cases, based on what was known about the virus at that time. Following the outbreak of SARS and MERS, WHO issued new guidance on how to prevent the spread of respiratory viruses recommending droplets and contact precautions for aerosol generating procedures conducted by health workers.

Approximately 215 countries and territories worldwide have been affected by the Covid-19 disease. According to WHO, it was reported that COVID-19 disease has caused 3,733,980 death cases world wild as of June 2020. United States of America hold the record with most number of death cases, which is 612,366 cases, followed by Brazil, which recorded to have 473,495 of death cases and India with 349,229 of death cases.

According to Malaysian new's agency, Bernama, Malaysia first reported its cases involving three Chinese tourists on 25th January 2020 who had enter Malaysia via Singapore on 23rd January 2020. The Ministry of Health (MOH) advised not to travel to China if there's no necessity. On 29th January, MOH confirmed additional three more cases, bringing cumulative total to seven positive cases. After series of cases reported outside China, Tedros Adhanom, the DirectorGeneral of WHO declared the novel coronavirus outbreak (Covid-19) a Public Health Emenrgency of International Concern (PHEIC).

According to one of the Malaysian new's agency, The Star, Malaysia recorded the first two deaths from Covid-19 on March 17. The Ex-Health Minister Datuk Seri Dr Adham Baba said the death involved case number 178, which was reported in Johor and case number 358, reported in Sarawak. According to Dr Adham, case 178 is a Malaysian aged 34, who had attended the tabligh gathering at Masjid Jamek Seri Petaling. Meanwhile, case 358 is a 60-year-old Malaysian who had a history of chronic illness (Mazwin, 2020).

1.2 Prediction of death due to COVID-19 using machine learning

Machine learning is a type of artificial intelligence (AI) that allows software applications to predict outcomes without being explicitly programmed to do so. Machine learning algorithm uses historical data as input to predict new output values.

Early prediction of patient mortality risks during a pandemic can decrease mortality by assuring efficient resource allocation and treatment planning. The data of new cases, death cases and recovered cases can be used to predict the probability of death due to COVID-19 disease. The factor affected the increasing in mortality rate are still being study. So, it is very important to identify which factors most affected the increased in number of deaths in one country.

Health workers are often unable to accurately predict the prognosis of COVID-19 patients upon their admission until later stages of the disease. Furthermore, the course of COVID-19 can take unpredictable turns where the condition of a seemingly stable patient deteriorates rapidly to a critical state. To enhance clinical prediction, AI models could be valuable assistants since they can detect complex patterns in large datasets. The predicted result can be used by the health worker to plan and predict which patients require more attention regarding their condition and health.

1.3 Problem Statement of the project

When employing a prediction algorithm, the result may differ depending on the parameter used. In this study, only one prediction model was used, which is the ARIMAX model. The problems are listed as below :

1. Based on findings, there is no previous study was identified on investigation on the prediction of death of COVID-19 focusing on state or area in one country.
2. The correlation between the time series data for each statess, patient's age and medical history has not been studied in previous works.
3. There is limited knowledge about the relationship between the number of deaths in one day with the number of new confirmed cases from yesterdays' according to the lag of days in data set.

1.4 Objectives of the project

The main goal is to determine which parameters give the best prediction result to forecast the probability of death due to COVID-19 disease. To achieve the aim of this project, the objectives are listed as follows:

1. To predict the probability of death due to COVID-19 disease based on previous time series data of death cases and new confirmed cases.
2. To compare the result based on different prediction input parameters, which are state, patients' age and medical history.
3. To investigate the relationship between the number of death in one day with the previous number of cases up to 10 days.

1.5 Thesis overview

This thesis starts with chapter 1 which explains the background of the projects and the importance of the project, followed by chapter 2 which discussed the literature review of the past findings. Then, continue with the method and algorithm of the prediction model used in this project in chapter 3. Then, results and discussion were discussed in chapter 4 and finally the summary and future work on the finding is discussed in chapter 5.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The previous studies of the prediction of death of COVID-19 using machine learning are presented and reviewed in this subchapter. The chapter starts with subchapter 2.2 which discusses the previous findings on COVID-19 outbreaks. Followed by section 2.3, previous works on predicting the COVID-19 confirmed cases and death cases using machine learning were discussed. Next, ARIMA and ARIMAX time series model for predicting COVID-19 death cases from previous findings are also discussed in the last sections.

2.2 Early findings on COVID-19 outbreak

In late December 2019, the world was shocked with a new unknown disease which was caused by a virus named Coronaviruses. The disease was believed to start in Wuhan, China. After several findings, the disease was named after COVID-19 which stands for Corona Virus Disease. Several clusters of cases then were reported outside Wuhan which led to a pandemic outbreak. WHO has published several measures to prevent the outbreak. The disease was believed to be transmitted via droplets and close contacts between humans. Li Q et al., 2020 found that there is evidence that human-to-human transmission has occurred among close contacts. Based on the data collected, the basic reproductive number (R_0) was believed to be 2.2. The infected person can infect the other 2.2 person via close contact. They also add that the number of infected person will increase as long as the value of R_0 is greater than 1. Several measures need to be taken seriously in order to reduce the value of R_0 to less than 1.

The number of infected cases keep on increasing until it involved almost all countries in the world. Malaysia recorded the first confirmed case on 25th January 2020 involving three Chinese tourists. The infected patients were observed to have several symptoms such as fevers, dry cough and tiredness. Elengoe et al., 2020 also adds that some patients may suffer from severe pneumonia, organ failure, acute respiratory tract infection and septic shock, which can lead to death. The authors also add that there are some infected individuals who do not develop any symptoms and do not feel unwell. These people are called asymptomatic carriers. From the study, patients who are highly at risk to COVID-19 are the elderly, young children, pregnant ladies, and people with chronic diseases such as hypertension, diabetes, heart problems, kidney and liver diseases.

2.3 Previous work on predicting COVID-19 confirmed cases and death cases using machine learning

As mentioned in previous chapter, machine learning algorithm uses historical data as input to predict new output values. An early study from Malaysia has created a representation model for predicting the number of COVID-19 cases using different neural network models, specifically Neural Network Auto-Regressive, Multi-Layer Perceptron, and Extreme Learning Machine (Purwandari et al., 2020).

The confirmed, recovered and death cases data were collected from 22nd January 2020 until 13th June 2020 with total of 143 days. The data then were separated into training and testing. The data from 7th June 2020 until 13th June 2020 were selected as testing data while the remaining data were selected as the training. This study used four main algorithms which are the Artificial Neural Network (ANN), Multi-Layer Perceptron (MLP), Neural Network Auto Regression Model (NNAR) and extreme Learning. Based on the result, the models were compared and the forecast is

made for the next 7 days. In conclusion, this study has indicated that the MLP model is the best model for 7-step ahead forecasting for confirmed, recovered, and death cases in Malaysia. However, according to the result of testing data, the ELM performs better than the MLP model.

Another study from Malaysia found that Singular Spectrum Analysis Recurrent Forecasting (SSA-RF) would discriminate noise in a time series trend and produce significant forecasting results (Shaharudin et al., 2020). The study aim to forecast the COVID-19 cases for the health authorities to comprehend the outbreak pattern better, and take further actions to stop the epidemic transmission. The data were collected from 25th January 2020 until 29th April 2020 with total of 95 days. First, the data were decomposed into components which could be defined in the forms of trend, seasonal, and noise components by using the Singular Spectrum Analysis (SSA). The results showed that parameter was suitable for short series outbreak data and the appropriate number of leading ET s to obtain was crucial as it affected the forecasting outcomes.

According to Li et al., with appropriate data set, hospitals can heuristically predict whether or not a patient requires immediate care (Li et al., 2020). This can lessen the fatality rate of patients infected from COVID-19. However, it is difficult to predict the high fatality risk of a patient who should be admitted to a hospital with high priority since there are different factors that contribute to an individual's infection progression once they were confirmed positive for covid-19. The data of the patients' age, sex, city, province, country date admission and date confirmation to hospitals were recorded from 14th May 2020 until 21st June 2020.

Two different datasets were used, one that specified symptoms and comorbidities and one that generalized across them. The data was separated into 70% for training and 30% for testing and validation. The results using Logistic Regression, SVM, Random Forest, One-class SVM, Isolation Forest, local outlier factor (lof) and autoencoder were evaluated with multiple metrics, including accuracy, specificity, sensitivity and the area under the curve (AUC) score. Overall, the best results were obtained by the auto encoder model and the matrix display a strong correlation between fatality and COVID-19 patients with chronic diseases.

Table 2.1 Summary of Previous work on predicting COVID-19 confirmed cases and death cases using machine learning

Author	Data size	Parameter	Algorithm	Result	Limitation
Purwandari et al., 2020	22/1/2020 - 13/6/2020	confirmed, recovered, and death cases of COVID-19	ANN MLP NNAR ELM	MLP model is the best model for 7-step ahead forecasting	Simple model cases without considering other factors that affect the number of these cases.
Shaharudin et al.,2020	25/1/2020 - 29/4/2020	Daily COVID-19 prevalence data	SSA-RF	F-SSA model suitable for short-time series outbreak data	Sudden spike in data leads to low performance of forecasting results.
da Silva et al.,2020	April 2020 - May 2020	meteorological data temperature, humidity, rainfall	EEMD ARIMA EEMD-ARIMAX	ARIMAX RMSE - 211.987 standard deviation -186.335 EEMD-ARIMAX RMSE - 155.330 standard deviation - 145.645	-
Li et al.,2020	May 14 and June 21	Patients' age, sex, city, province, country, date admission, presence history of chronic illnesses.	SVM	Autoencoder sensitivity - 0.4.	Lack of abundant quality data used to train the models created.

2.4 ARIMA and ARIMAX time series model

AutoRegressive Integrated Moving Average with Exogeneous variables (ARIMAX) model are time series analysis methods, often used in statistical modeling to analyze changes that occur over time (Cao et al., 2020). ARIMAX model is basically an extended form of ARIMA model with an intervention sequence. ARIMA is suitable for all kinds of data, including non-stationary data, which is if there is no systematic change in mean, no trend, no systematic change in variance and periodic variations have removed (Shumway and Stoffer, 2010). Yang et al., 2020 add in order to build ARIMA for particular time series analysis, one should follow four phases which are Model identification, Estimation of model parameters, Diagnostic checking for the identified model and Application of the model.

2.5 ARIMAX model to forecast the number of death due to COVID-19

A study in China has collected the number of cases, number of suspected cases, the number of people in recovery, number of deaths and number of people in quarantine from 10th January 2020 until 9th February 2020 with total of 30 days. ARIMA and ARIMAX models in this study, the first-order difference (0,1,0) and the second-order difference (0,2,0) to smooth the original sequence According to the authors, the second order ARIMAX (0,2,0) models with the lowest value of R-squared 0.958 compared to the first-order difference can be used to predict the short-term development of the confirmed number of patients with COVID-19 outbreak.

Ajidobe and Adeboye, 2020 also add that ARIMAX with confirmed cases as exogenous variable model was found to outperformed ARIMA model. ARIMAX model was selected as the best disease predicting model in the study. The study extracted the data of death cases and confirmed cases in time frame of 43 weeks in Nigeria. From the evaluation, ARIMAX model has the lowest value of RMSE, 8.538211 and MAE, 5.9774 as compared to ARIMA model with RMSE value of 8.962229, and MAE value of 6.2152.

2.6 Summary on the study of previous works

Based on the discussed previous works, most of the prediction of death due to COVID-19 did not focus on the trend of confirmed cases and death cases in multiple areas or state in a country. Number of cases in each area plays a major role to indicate the number of death in certain country. Moreover, since there are no studies to date on investigating the relationship between the new confirmed cases and death cases for each state, patient's age and medical history, this project uses exogeneous factors data, such as the new confirmed cases and death cases data of each state, patient's age and medical history to predict the probability of death in Malaysia. Also, this study will be focus on the trend of death cases for each state, patients' age and medical history. This project will also investigate the correct time lag for the new confirmed cases with the actual death cases for each of the state and patients' age.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The method used in this project were discussed in this chapter. First, the flowchart of the method was explained in section 3.2, followed by the specification of the device used for data acquisition in section 3.3. Next, the data collection method, preprocessing steps, and data analysis were also discussed and recorded in section 3.4, 3.5 and 3.6 respectively. Lastly, the flowchart and algorithm of the model used for prediction in this project were explained in the last section.

3.2 Flowchart of the prediction process

This section briefly explained the steps for the prediction using machine learning. The process started with the data collection. All the data collected in this project are available via the official Malaysian Ministry of Health website. Once the data were collected, the preprocessing steps were done to generate a data set which can be easily read by the machine learning.

After that, the data were analyzed based on its category or subset. Data analysis is important in research because it makes studying data a lot simpler and more accurate. The trend of the new confirmed cases and death cases for each state, patients' age and medical history can be studied easily and precisely. Then, the data were divided into two sections training and testing before applying the selected machine learning model. However, there is only one time series model used in this project which will be discuss thoroughly in the next section.

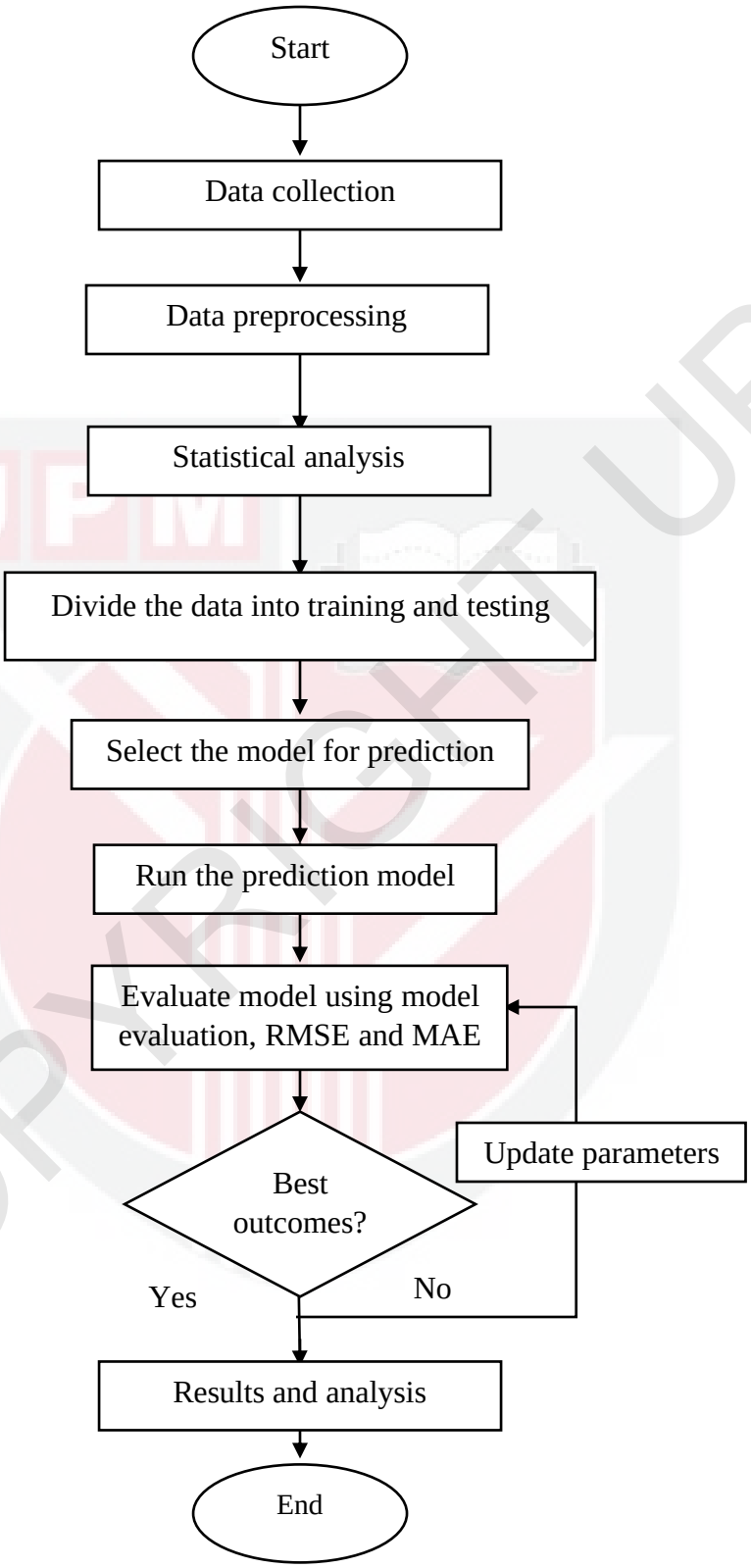


Figure 3.1 Flowchart of the model prediction process

3.3 Device and software used for data acquisition

The device used for this experiment is Asus brand A510U laptop. The computer specification was fundamental so that the software can run smoothly and efficiently. The following Table 3.1 describes the important features of Asus brand A510U laptop with the details.

3.1 Features of Asus brand A510U

Device name	LAPTOP-FDT3C351
Processor	AMD A6-9225 RADEON R4, 5 COMPUTE CORES 2C+3G 2.60 GHz
Installed RAM	4.00 GB
Edition	Windows 10
Version	21H2
System type	64-bit operating system, x64-based processor

The software used for analyzing data in this project is MATLAB ver. R2021a by MathWorks, Inc. MATLAB is used for matrix manipulations, plotting of functions and data, implementations of algorithms, creation of user interfaces which support multiple programming languages such as C++, Java, C, C#, Fortan or Phyton. In this research, MATLAB is used for plotting of graphs and development and implementation of algorithms. It is also used for organizing the data collection.

3.3 Data collection

The daily pandemic announcement provides basic data of each patient's background such as their age, sex, place of death and medical history. The data were obtained from The Ministry of Health (MOH) Malaysia official website, https://covid-19.moh.gov.my/terkini from 8th October 2020 to 13th July 2021, with total of 277 days. The data collected includes the cumulative number of confirmed cases and cumulative number of death cases for each state, patient's age and medical history.

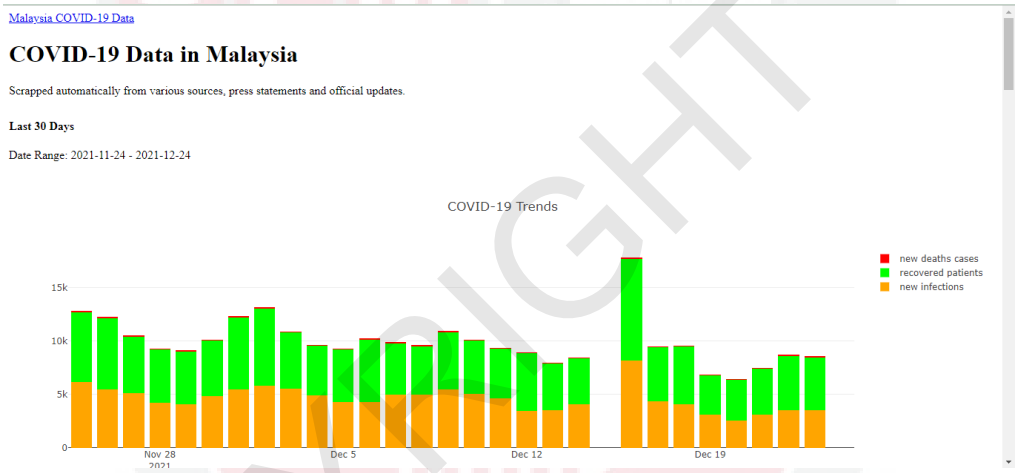


Figure 3.3 Front page of Ministry of Health (MOH) Malaysia website

(Source: MOH official website from https://covid-19.moh.gov.my/terkini)

3.4 Data preprocessing

Data preprocessing involves converting raw data into well-formatted data sets so that data mining analysis can be applied. The original data is usually incomplete and inconsistent in format. The data preprocessing were done to transform the raw data into an understandable format by the machine learning. Stated below are the figure of the raw data before and after undergo the preprocessing process and the flowchart of process.

The major difference between Figure 3.4 and Figure 3.5 is that the output data is well ordered, less volume and neat. The data were transformed into binary numbers so that the machine learning can trace and read. The major aim to transform the original format words into numbers so that it can be easy to be read by the machine learning. The unwanted data were being ignored to reduce the volume of the data so that it will take less time to work with the data. The flow of the process will be discussed thoroughly below.

	A	B	C	D	E	F	G	H	I
1	No. of death	State	Sex	Age	Place of death	Diseases	Date	No. of cas	Nationality
2	142	Sabah	M	55	Hospital Lahad Datu	Diabetes; Hypertensive heart disease; Dyslipidemia.	8/10/2020	12489	Citizens
3	143	Sabah	F	75	Hospital Semporna	Hypertensive heart disease.	8/10/2020	13692	Citizens
4	144	Sabah	M	57	Hospital Duchess of Kent Sandakan	Diabetes; Hypertensive heart disease.	8/10/2020	13705	Citizens
5	145	Sabah	M	82	Hospital Tawau	N/A	8/10/2020	14261	Citizens
6	146	Sabah	M	53	Hospital Tawau	Diabetes	8/10/2020	14264	Citizens
7	147	Sabah	M	54	Hospital Tawau	Hypertensive heart disease; Asthma	9/10/2020	12991	Citizens
8	148	Sabah	M	68	Hospital Semporna	N/A	9/10/2020	14028	Citizens
9	149	Sabah	F	66	Hospital Queen Elizabeth	Diabetes; Hypertensive heart disease; Dry cough	9/10/2020	14151	Citizens
10	150	Sabah	F	58	Hospital Tawau	N/A	9/10/2020	14640	Citizens
11	151	Sabah	F	57	Hospital Semporna	Hypertensive heart disease	9/10/2020	14641	Citizens
12	152	Sabah	M	64	Hospital Queen Elizabeth	Benign prostatic hyperplasia (BPH)	9/10/2020	14642	Citizens
13	153	Sabah	M	61	Hospital Tawau	Diabetes; Tuberculosis	10/10/2020	14038	Citizens
14	154	Sabah	F	54	Hospital Queen Elizabeth	Diabetes; Hypertensive heart disease; Obesity	10/10/2020	14974	Citizens
15	155	Sabah	M	51	Hospital Semporna	Diabetes; Hypertensive heart disease; Heart disease	10/10/2020	14999	Citizens
16	156	Sabah	M	67	Hospital Duchess of Kent	Tuberculosis (TB); Heart disease	11/10/2020	14259	Citizens

Figure 3.4 The raw data before employing preprocessing process

	A	B	C	D	E	F
1	No. of dea	State	Sex	Age	Diseases	Date
2	142	12	1	55	23; 33; 26	10/8/2020
3	143	12	2	75	33	10/8/2020
4	144	12	1	57	23; 33	10/8/2020
5	145	12	1	82	72	10/8/2020
6	146	12	1	53	23	10/8/2020
7	147	12	1	54	33; 6	10/9/2020
8	148	12	1	68	72	10/9/2020
9	149	12	2	66	23; 33; 25	10/9/2020
10	150	12	2	58	72	10/9/2020
11	151	12	2	57	33	10/9/2020
12	152	12	1	64	9	10/9/2020
13	153	12	1	61	23; 61	10/10/2020
14	154	12	2	54	23; 33; 43	10/10/2020
15	155	12	1	51	23; 33; 30	10/10/2020
16	156	12	1	67	61; 30	11/10/2020

Figure 3.5 Data after employing the preprocessing process

Based on the Figure 3.4, raw data are incomplete, large volume and have inconsistent languages with lots of typing errors. The first step of data preprocessing is to decide which data are required for data cleaning and apply data cleaning technique to the incomplete data. The process for data cleaning includes filling in missing values or deleting rows with missing data, smoothing the noisy or resolving the inconsistence such as the language, spelling, or format of the data. There are four missing data were being ignored because it doesn't affect the large size of data. Data inconsistencies can occur due to human errors especially during storing data process.

Once the data were cleansed, the next step is to check the volume of data and identify which technique is suitable in order to reduce volume of data. The volume of data is huge, data reduction aims simplified representation of the data. For example, the raw data includes the data of the place of death, and the patients' nationality which was not needed in this project. The data were deleted to reduce the size of the data and make it well ordered.

The last step is to undergo the data sampling process. This is due to time, storage or memory constraints, a dataset is too big or too complex to be worked with. Sampling techniques can be used to select and work with just a subset of the dataset, if it has approximately the same properties of the original one. The data were arranged in subset, making it easier to be analyzed. In result, output data become reliable, accurate and less volume. The flowchart of the preprocessing were shown in Figure 3.5.

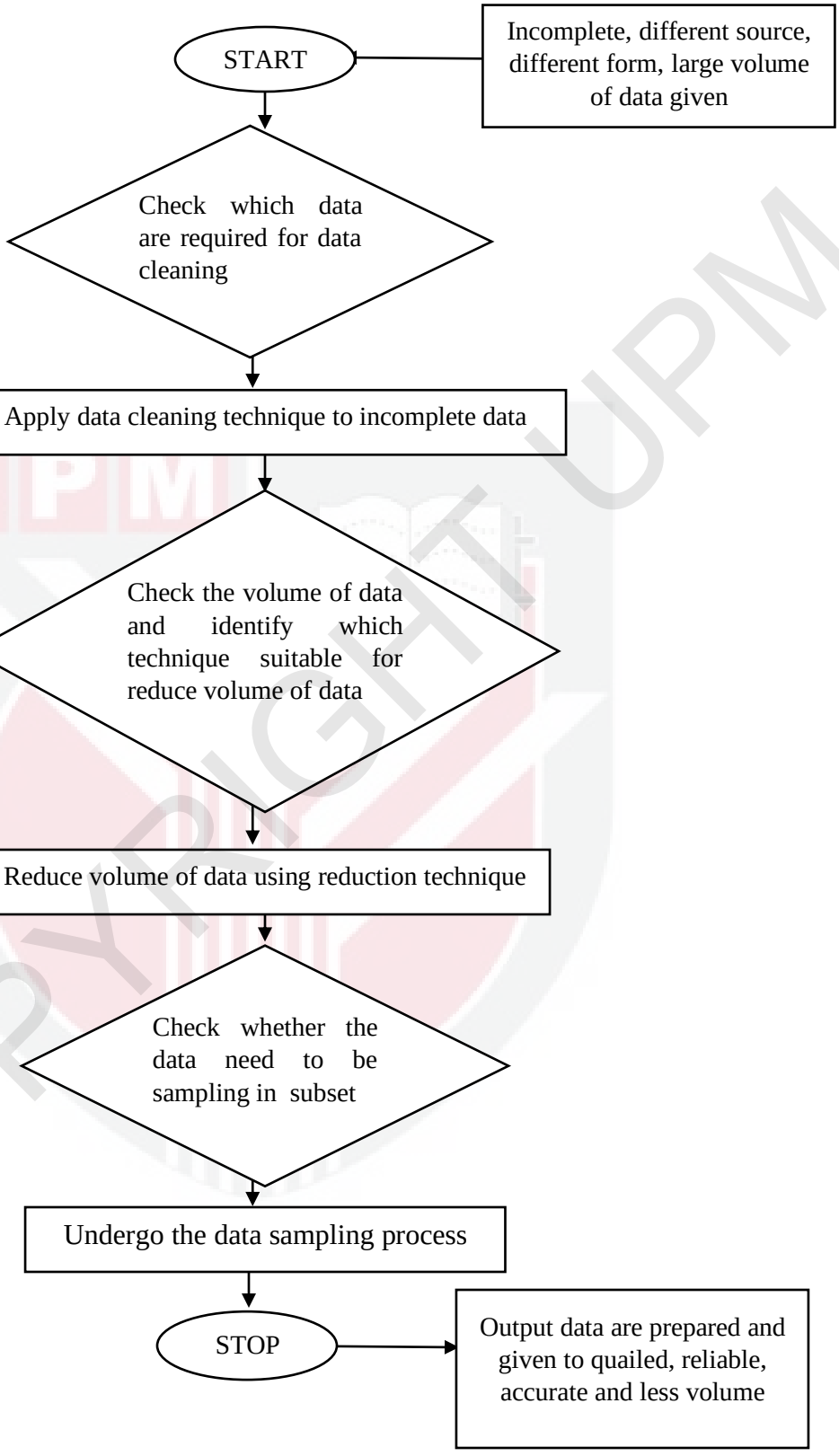


Figure 3.6 Flowchart of data preprocessing

3.5 Statistical analysis

Total number of positive cases in Malaysia from 8th October 2020 until 13th July 2021 is 832,083 with total of 6,173 of death cases, which makes the number of survived cases is 825,910. The data of positive cases and death due to COVID-19 disease were analyzed by month, state, and patients' details which include their age, sex and their medical history.

From figure 3.7, the number of positive cases was at it's peak on June 2021. There is an increment in number of positive cases on January 2020 but continue to decrease until March 202. The number of cases later started to increase drastically in April 2020 until June 2021. Meanwhile, from figure 3.7, the number of death due to Covid-19 disease was at it's peak on July 2021. It has a slight increment in February 2020 and decrease until March 2021. The number of death cases later started to increase on April 2021 and continue to increase drastically from May 2021 until June 2021. The trend of new cases effected the trend of the mortality rate. As the number of cases increases, the number of death will also increases, and vice versa.

From table 3.2, the mortality rate is the probability of one person died from the total confirmed cases reported. The highest value of mortality rate was on June 2021 with 0.01168 compared to other months. The total mortality rate from October 8th, 2021 to July 13, 2020 is 0.0587.

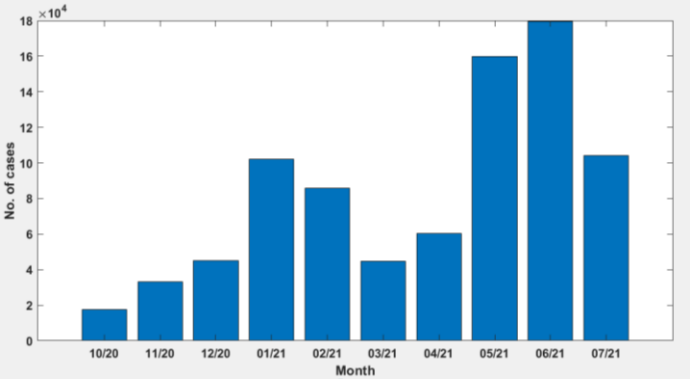


Figure 3.7 No. of positive cases from October 2020 to July 2021

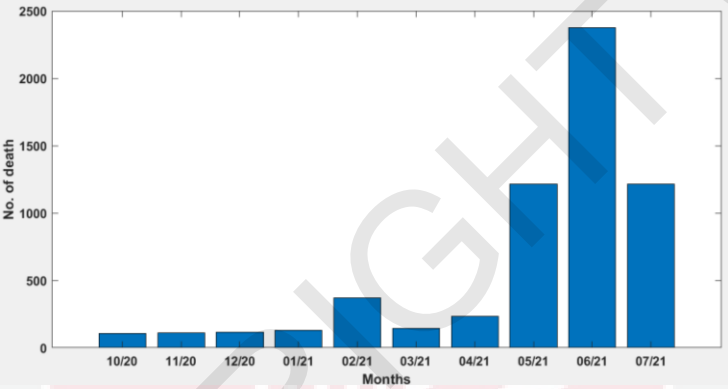


Figure 3.8 No. of death due to COVID-19 disease from October 2020 to July 2021

Table 3.2 Mortality rate of COVID-19 disease from October 2020 to July 2021

Year	Month	Cases	Death	Survived	Mortality rate
2020	October	17,555	108	17,447	0.0061
	November	33,117	114	33,003	0.0034
	December	45,079	112	44,967	0.0024
2021	January	101,949	289	101,660	0.0028
	February	85,793	370	85,423	0.0043
	March	44,748	142	44,606	0.0031
	April	60,338	234	60,104	0.0038
	May	159,911	1,214	158,697	0.0075
	June	179,622	2,374	177,248	0.0132
	July	103,970	1,215	102,755	0.0116
Total		832,082	6,172	825,910	0.0587

Based on Figure 3.9, for each state in Malaysia, Selangor has the highest number of death while Perlis has the least number of death due to the COVID-19 disease outbreak. Second highest is Johor followed by Sabah compared to other state. The number of positive cases has affected the death rate for each state. The higher the positive cases for each state, the higher the number of death cases in each state, and vice versa.

The data of death for each state were then divided into two, for training and testing for prediction. For each state, the training data were selected from 8th October 2020 until 31st May 2021. The testing data were selected from 1st June 2021 until 13th July 2021. The total training and testing data differs for each state depends on the number of death in each state. From the prediction result, the lagging data of 10 days were then observed.

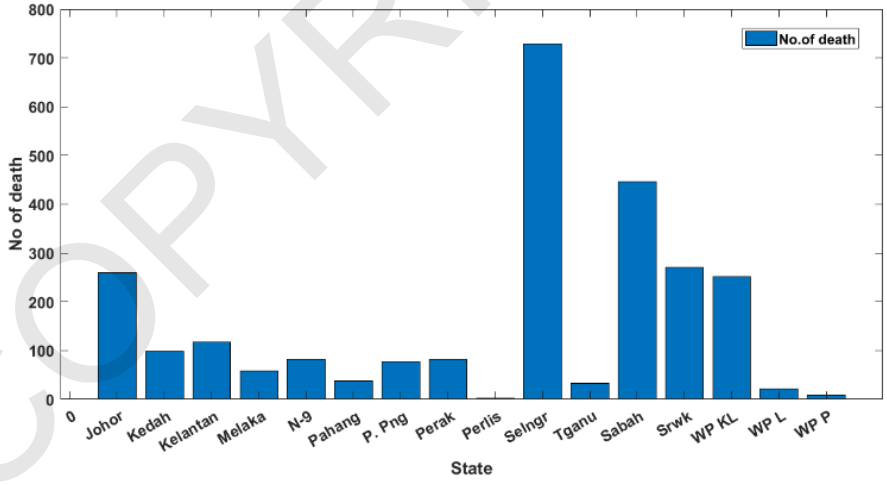


Figure 3.9 No. of death due to COVID-19 disease in each state

Based on the data collected, the percentage of male patients died from COVID-19 disease are higher than female with 59% to 41% as shown in Figure 3.10 below. A recent study had compared differences in immune responses between male and female COVID-19 patients and found that males had higher circulating innate inflammatory cytokines, interleukin (IL)-8 and IL-18, and stronger induction of non-classical monocytes while females had more robust T cell activation and higher interferon (INF)α2 (Takahashi et al, 2020). In research from Kadel and Kovats, 2018 these differences in immune response may explain the different COVID-19 incidence and clinical outcomes between sexes. Estrogens in female have been reported to have immunostimulatory or immunosuppressive effects depending on concentration and cell types, while testosterone in male is immunosuppressive, which in short, male has less defensive immunity to fight new viruses infection compared to female.

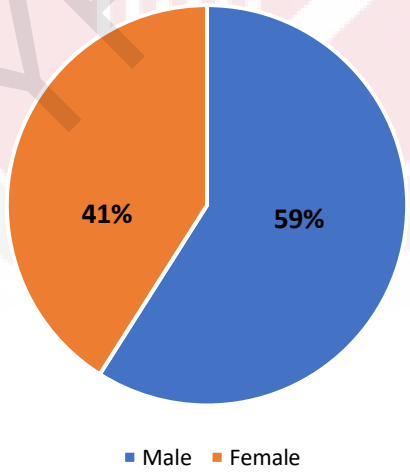


Figure 3.10 Percentage of male and female patients died from COVID-19 disease

The data of total death due to COVID-19 disease were then being analyzed by patients' age. Since there is total of 104 unique ages, the data then were determined into class intervals of five as shown in Table 3.3 below.

Based on the Table 3.2, the mode class is 60-64 years old. Patients who died from COVID-19 disease are mostly in between age 60 – 64 years old. While the mean age is 64 years old, which means average age of patients who died from COVID-19 is 64 years old. Meanwhile patients who are from 5 to 14 years old have the least number of death. From Figure 3.8, the shape of the histogram is bell shaped with skewed to the left which indicates that the data are normally distributed and majority of patients died are above 60 years old.

From the class intervals, the data were divided into two parts, training and testing for prediction. The training data were selected from 8th October 2020 until 30th May 2021. The testing data were selected from 1st June until 13th July 2021. Total training and testing data differs for each class depends on the total number of death for each class.

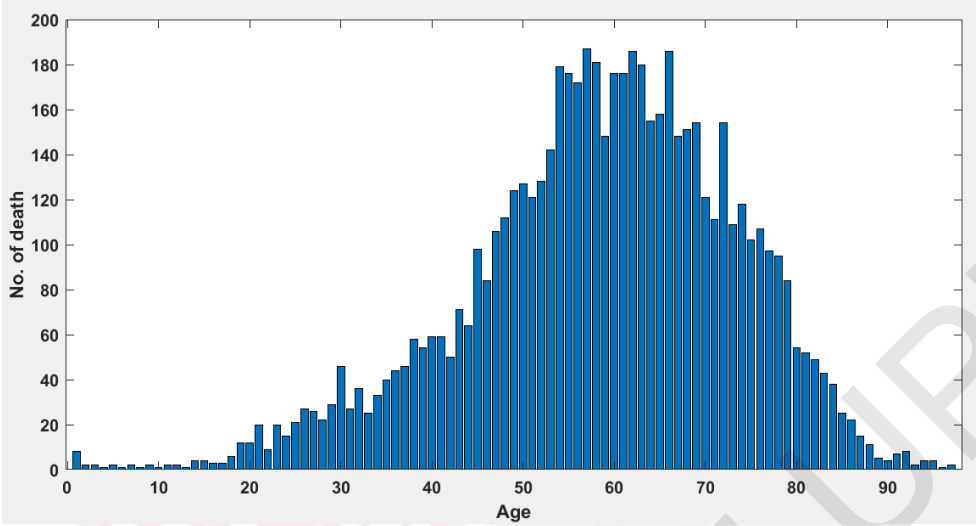


Figure 3.11 Number of patients died from COVID-19 disease by age group

Table 3.3 Number of patients died from COVID-19 by age group

Age	No. of Death
0-4	13
5-9	3
10-14	3
15-19	8
20-24	20
25-29	73
30-34	111
35-39	163
40-44	221
45-49	293
50-54	464
55-59	642
60-64	895
65-69	866
70-74	798
75-79	649
80-84	519
85-89	382
90-94	114
95-99	26
100-104	11

Based on the collected patients' medical history data, there are total of 70 different background diseases of patients died from COVID-19 disease recorded. From 70 unique diseases, there are five diseases that contributes the most in the mortality rate of COVID-19 disease.

Table 3.4 shows five background diseases with highest number of patients died from Covid-19. Based on the table, 28.5% patients who died from COVID-19 suffered from hypertensive heart disease or also known as high blood pressure. Followed by Diabetes (22.2%), Kidney disease (6.3%), Heart disease (8.3%) and Dyslipidemia (8.7%). Only 15.1% patients died from COVID-19 doesn't have any disease in their medical record. By mean, patients who suffered from critical diseases has 84.9% chances to die from COVID-19 disease once they're infected.

Based on the record of the five mentioned diseases, the data were divided into two, training and testing data for prediction. The training data were selected from 8th October 2020 until 31st May 2021. While the testing data were selected from 1st June 2021 until 13th July 2021. Total of training and testing data differs for each diseases depends on the number of death.

Table 3.4 Five background disease of patients died from COVID-19 disease

Diseases	Total death	Probability
Hypertensive heart disease	3,988	0.285
Diabetes	3,107	0.222
Kidney disease	8,78	0.063
Heart disease	1,160	0.083
Dyslipidemia	1,217	0.087

3.6 Model used for prediction

This project mainly generates a statistical model based on time series analysis. The collected data is a large sample data, the model that is used in this project is suitable for containing parameters to be estimated to predict the trend of the COVID-19 based on time series analysis. The model that were used in this project is Auto Regressive Integrated Moving Average with Exogenous (ARIMAX).

The standard autoregressive integrated moving average (ARIMA) model allows to make forecasts based only on the past values of the forecast variable. The model assumes that future values of a variable linearly depend on its past values, as well as on the values of past stochastic shocks. The ARIMAX model is an extended version of the ARIMA model. It includes also other independent predictor variables. The model is also referred to as the vector ARIMA or the dynamic regression model. The ARIMAX model is similar to a multivariate regression model but allows to take advantage of autocorrelation that may be present in residuals of the regression to improve the accuracy of a forecast.

3.7 ARIMAX algorithm

ARIMA model is a generalization of an autoregressive moving average (ARMA) model. The autoregressive moving average model including exogenous covariates, ARMAX(p,q), extends the ARMA(p,q) model by including the linear effect that one or more exogenous series has on the stationary response series y_t. The general form of the ARMAX(p,q) model is

$$y_t = \sum_{i=1}^{p} \emptyset_i y_{t-i} + \sum_{k=1}^{r} \beta_k x_{tk} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}, \quad (1)$$

and it has the following condensed form in lag operator notation:

$$\phi(L)y_t = c + x'_t\beta + \theta(L)\varepsilon_t \quad (2)$$

In equation (2) the vector *x't* holds the values of the *r* exogenous, time-varying predictors at time *t*, with coefficients denoted *β*. This model can be used to check if a set of exogenous variables has an effect on a linear time series.

ARMAX models have the same stationarity requirements as ARMA models. Specifically, the response series is *stable* if the roots of the homogeneous characteristic equation of

$$\phi(L) = L^P - \phi_1 L^{P-1} - \phi_2 L^{P-2} - \ldots - \phi_P L^P = 0 \quad (3)$$

lie outside of the unit. In equation (1) if the response series y_t is not stable, then it can be difference it to form a stationary ARIMA model. This can be done by specifying the degrees of integration D. When specify an AR model using ARIMA, the software displays an error if coefficients that do not correspond to a stable polynomial being enter. Similarly, estimate imposes stationarity constraints during estimation.

The software differences the response series y_t before including the exogenous covariates if you specify the degree of integration D. In other words, the exogenous covariates enter a model with a stationary response.

Therefore, the ARIMAX(*p*,*D*,*q*) model is

$$\phi(L)y_t = c^* + x_{,t}\beta + \theta^*(L)\varepsilon_t \quad (4)$$

where $c^* = c/(1-L)^D$ and $\theta^*(L) = \theta(L)/(1-L)^D$. Subsequently, the interpretation of *β* has changed to the expected effect a unit increase in the predictor has on the *difference* between current and lagged values of the response. The software treats the exogenous covariates as fixed during estimation and inference.

3.8 Model evaluation

3.8.1 Root Mean Square Error

Root Mean Square Error is the measure of how well a regression line fits the data points. RMSE can also be construed as Standard Deviation in the residuals. It is a standard way to measure the error of a model in predicting quantitative data, (Moody, 2021). Formally it is defined as follows

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{5}$$

Based on equation 5, $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are the predicted values, $y_1, y_2, \ldots, y_n$ are observed values and n is the number of observations. By ignoring the division by n under the square root, resemblance to the formula for the Euclidean distance between two vectors in $\mathbb{R}^n$:

$$\text{distance}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{6}$$

RMSE can be thought of as normalized distance between the vector of predicted values and the vector of observed values. The observed values are determined by adding random "errors" to each of the predicted values, as follows:

$$y_i = y_i + \epsilon_i \; for \; i = 1, \ldots, n \tag{7}$$

In equation (7), $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed errors. These errors, thought of as random variables, might have Gaussian distribution with mean μ and standard deviation σ, but any other distribution with a square-integrable PDF (probability density function) would also work.

The mean μ of the distribution of the errors would correspond to a persistent bias coming from mis-calibration, while the standard deviation σ would correspond to the amount of measurement noise. It can be seen through a bit of calculation that, the mean μ of the distribution for errors were determined, and would like to estimate the standard deviation σ. The equation is as follows,

$$\mathbb{E}\left[\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}\right]$$

$$= \mathbb{E}\frac{[\sum_{i=1}^{n} \epsilon_i^2]}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\,[\epsilon_i^2]$$

$$= \mathbb{E}[\epsilon^2]$$

$$= Var(\epsilon) + \mathbb{E}[\epsilon]^2$$

$$= \sigma^2 + \mu^2 \qquad (8)$$

$\mathbb{E}[\ldots]$ is the expectation, and $Var(\ldots)$ is the variance. Replace the average of the expectations $\mathbb{E}[\varepsilon_i{}^2]$ on the third line with the $\mathbb{E}[\varepsilon^2]$ on the fourth line where ε is a variable with the same distribution as each of the ε_i, because the errors ε_i are identically distributed, and thus their squares all have the same expectation.

Assumed that μ were already determined, that is, the persistent bias in the instruments is a known bias, rather than an unknown bias. It might as well correct for this bias right off the bat by subtracting μ from all our raw observations. That is, it might as well suppose the errors are already distributed with mean $\mu = 0$. Plugging this into the equation above and taking the square root of both sides then yields:

$$\sqrt{\mathbb{E}\left[\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}\right]} = \sqrt{\sigma^2 + 0^2} = \sigma \quad (9)$$

If removed the expectation $\mathbb{E}[\ ...\ ]$ from inside the square root, it is exactly the formula for RMSE form before. The central limit theorem tells us that as n gets larger, the variance of the quantity $\Sigma_i\,(\hat{y}_i - y_i)^2 / n = \Sigma_i\,(\varepsilon_i)^2 / n$ should converge to zero. In fact a sharper form of the central limit theorem tell us its variance should converge to 0 asymptotically like $\frac{1}{n}$. But then RMSE is a good estimator for the standard deviation σ of the distribution of our errors

To sum up, RMSE is a good measure to estimate the standard deviation σ of a typical observed value from the model's prediction, assuming that the observed data can be decomposed as, observed value = predicted value + predictably distributed random noise with mean zero.

If the noise is small, as estimated by RMSE, this generally means the model is good at predicting the observed data, and if RMSE is large, this generally means the model is failing to account for important features underlying the data.

3.8.2 Mean Absolute Error

In Machine Learning, MAE is a model evaluation metric often used with regression models. It can be said that MAE is, Prediction error = Actual value – Predicted value. This prediction error is taking for each record after which all error were convert to positive. This is achieved by taking Absolute value for each error as below,

Absolute Error → |Prediction Error|

Finally the mean were calculated for all recorded absolute errors. Mean average error equation:

$$\text{MAE} = \frac{\sum_{i=1}^{n} abs(y_i - \lambda(x_i))}{n} \quad (10)$$

Given any test data-set, MAE of one model refers to the mean of the absolute values of each prediction error on all instances of the test data-set. Prediction error is the difference between the actual value and the predicted value for that instance. Statistically, Mean Absolute Error (MAE) refers to the results of measuring the difference between two continuous variables.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

In this chapter, the cross correlation of the data of new confirmed cases and death cases with the data set are discussed in section 4.2. Next, the result of the prediction of probability of death for each state, patients' age and background diseases with lags of 10 days using the prediction model ARIMAX are recorded and discussed in section 4.3, 4.4 and 4.5 respectively.

4.2 Cross correlation of new cases and death cases data with the data set

In this project, three factors were studied to predict the probability of death due to COVID-19 disease. In this section, the correlation coefficient (cc) values were used to examined how similar the measurements of the data of new confirmed cases and death cases are across the dataset. When the value obtains near to +1 or-1, it represents that the input variable correlates the most to forecast the probability of death due to COVID-19 disease.

Figures shows the value of cc for the data of new cases and death cases with the data of death in each state, patient's age and background diseases. The indicators in the right in each of the figures indicates that the more the data correlates, the darker the blue color would be, and vice versa.

Figure 4.1 shows the cross correlation between the data of new cases and death cases with the data of death in each state. The average value of cc for the data of new cases with the data of death in each state is 0.6812, while the data of death cases is 0.5888. Overall, the data of new confirmed cases correlates more with the data of death in each state compared to the data of death cases. This means, the results of prediction of death for each state are more precise if the data of new cases were used as parameter compared to the data of death cases. Selangor has recorded the highest value of cc with 0.9522 while Sabah has recorded the least number of cc with -0.1303.

Meanwhile, in Figure 4.2, the average value of cc for the data of new confirmed cases is 0.6547 while the data of death cases is 0.7522. The data of death cases correlates more with the data of death for each age class compared to data of new cases. This means, the results of prediction of death for the patients' age are more precise when the data of death cases were used as input compared to the data of new cases. The data of patients with age 60-64 has recorded the highest value of cc which is 0.9371.

In Figure 4.3, the average value of cc for the data of new confirmed cases is 0.5788 while the data of death cases is 0.6491, which indicates that the data of death cases correlates more with the data of death for each of the patients' diseases compared to the data of new cases. This means, prediction of death for each of the diseases gives better result when the data of death per day were used as input compared to the data new cases.

The difference value of cc might be affected by the size of the training and testing data. Overall, the bigger the size of training and testing data, the more the data would correlate with each other. For example, Selangor has the highest number of death compared to other state. So, the value of cc for Selangor is highest compared to other states.

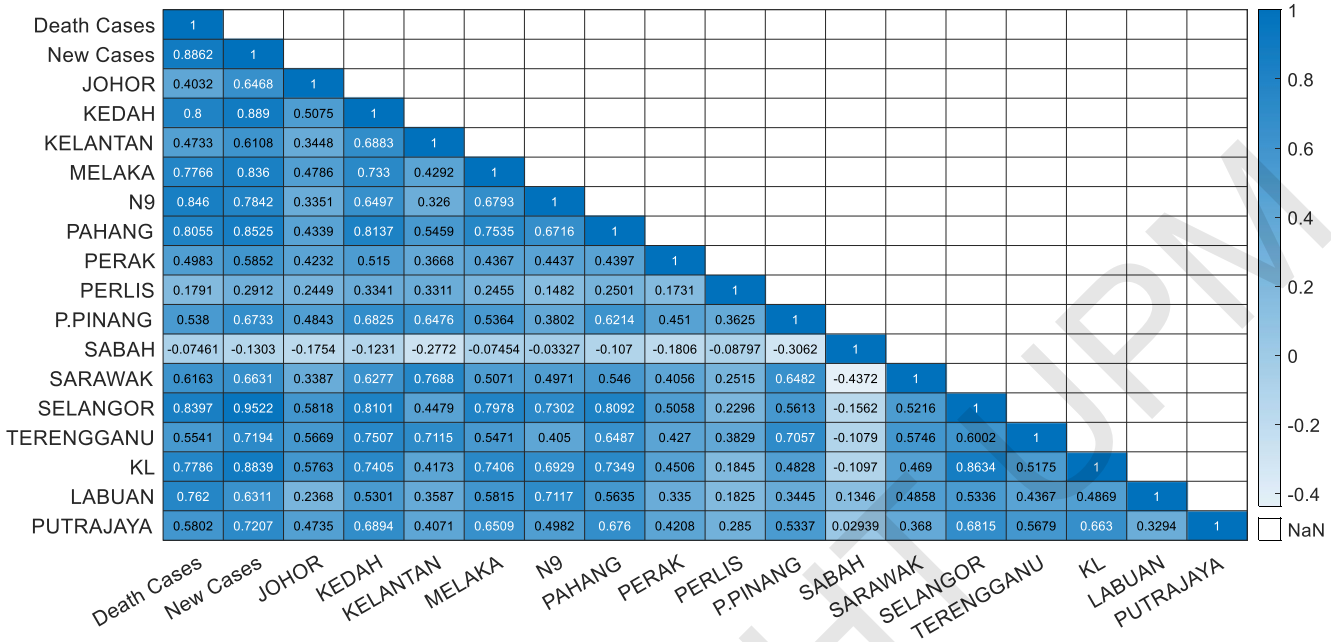


Figure 4.1 Correlation coefficient of death cases and new confirm cases with every state

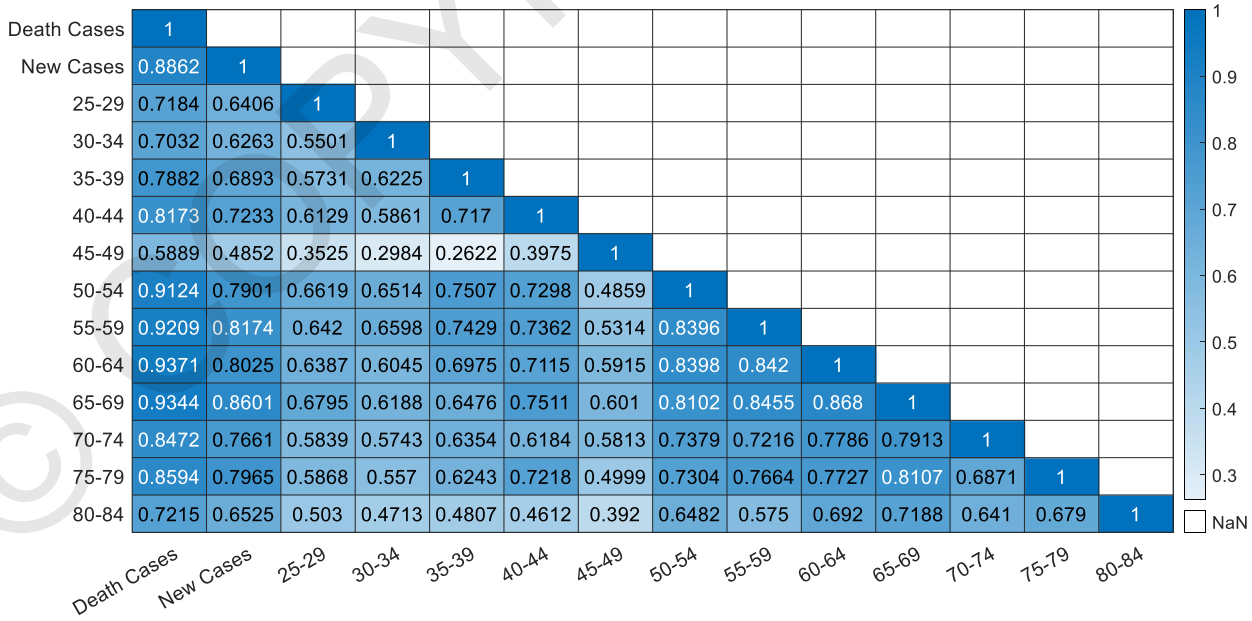


Figure 4.2 Correlation coefficient of death cases and new confirmed cases with patient's age

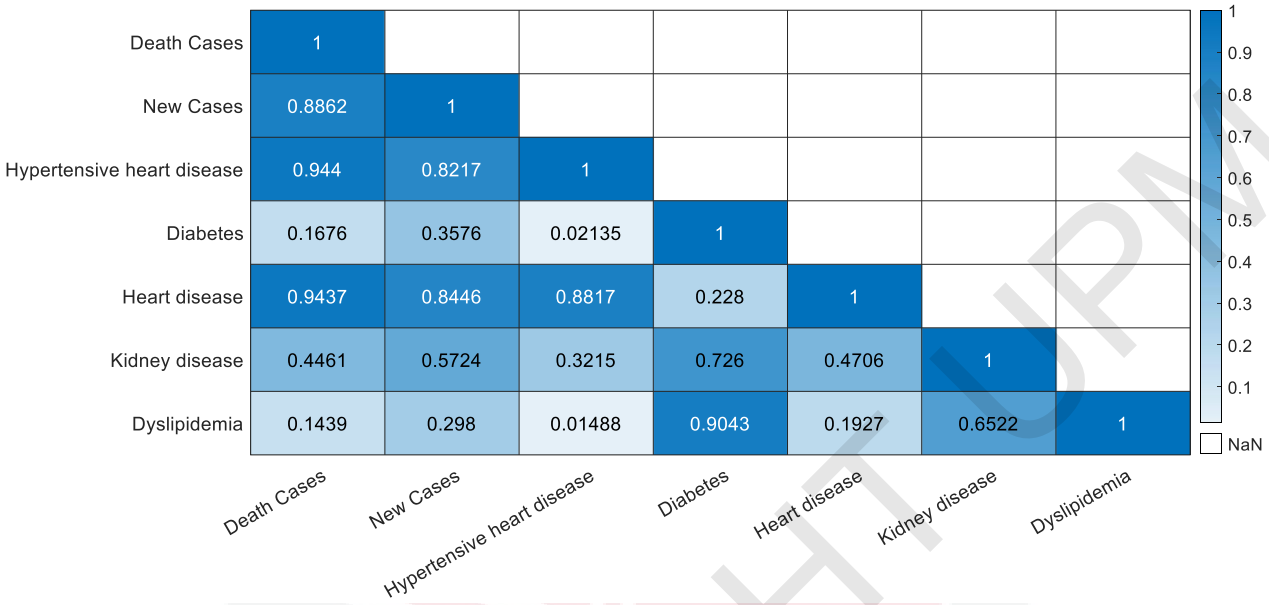


Figure 4.3 Correlation coefficient of death cases and new positive cases with patient's background diseases

4.3 Prediction of probability of death for each state

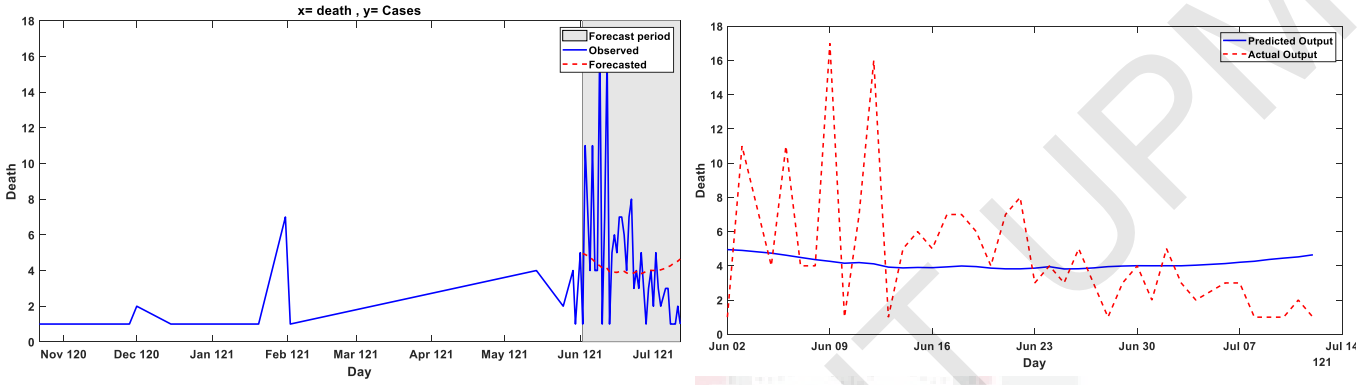
In this sub chapter, the data of new cases and data of death cases were used as parameters to predict the probability of death due to COVID-19 diseases in each state. The data of new cases gave the best result for prediction of death due to COVID-19 disease in each state compared to the data of death cases. Figure 4.4 (a) and (b) is the predicted result of death due to COVID-19 disease for W.P. Labuan and 4.5 (a) and (b) is the predicted result of death due to COVID-19 disease for Selangor. Both predicted results shown below used the data of new cases as the parameter.

As shown in each figure, the blue line in graph (a) indicates the observed data, which is the testing data while the red dotted line indicates the forecast data. The darker or gray area is basically the forecast period. For this case, the data from 1st June to 13th July are being forecasted using the testing data. On the other hand, for both figure (b), the blue line in graph indicates the predicted output and the red dotted line indicates the actual output.

The best predicted results are achieved when the value of the error is low and the value of correlation coefficient (cc) is high. As mentioned before, if the parameters are alike, the value of cc will be close to 1 and if they are entirely not alike then the correlation coefficient will be close to 0. This indicates that the higher the value of cc, the more the data of the parameters correlates with each other.

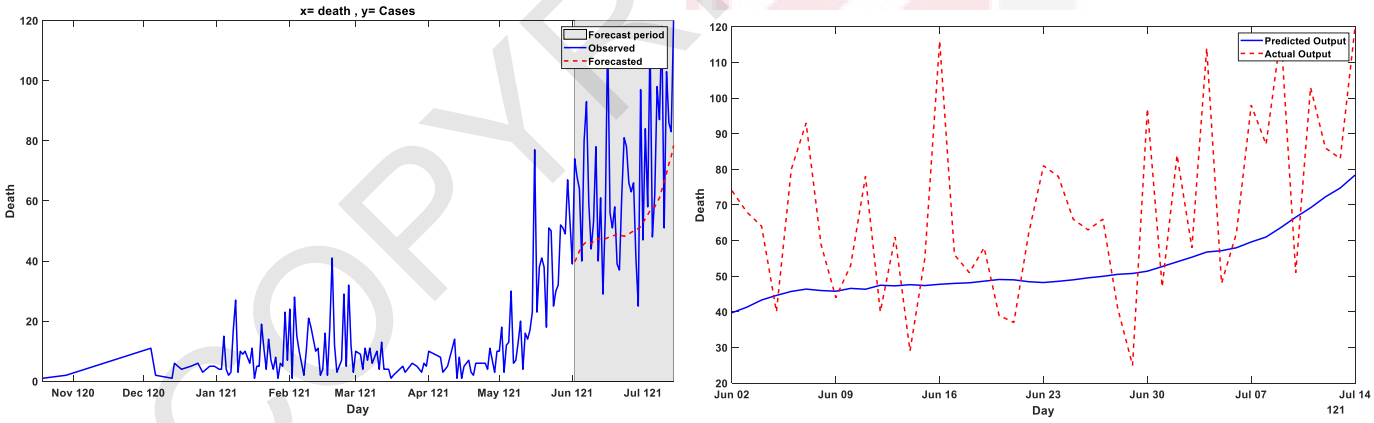
On the other hand, the lower the value of the error, the better the predicted result, which indicates that lower value of RMSE and MAE reflects the best result for prediction. Based on the Table 4.1, W.P. Labuan has recorded the least value of error, RMSE and MAE of 1.8696 and 1.4606 respectively with the highest value of cc, 0.7878 compared to the other state. The closer the value of cc to 1 indicates that the data of death in W.P. Labuan correlates with the parameter.

Meanwhile, Selangor has the highest error with the value of RMSE 36.8844 with cc of 0.4493. The evaluation of prediction of death due to COVID-19 disease for each state were shown in table 4.1 below.



(a) (b)

Figure 4.4 Prediction of death due to COVID-19 diseases in W.P. Labuan



(a) (b)

Figure 4.5 Prediction of death due to COVID-19 disease in Selangor

Table 4.1 Evaluation of prediction of probability of death due to COVID-19 disease in each state

State	Parameters					
	Death cases			New cases		
	cc	RMSE	MAE	cc	RMSE	MAE
JOHOR	0.1546	10.8864	9.0212	0.5886	17.3940	15.2983
KEDAH	0.5232	10.2709	9.2356	0.2937	4.8993	4.2694
KELANTAN	0.3925	8.9692	6.4711	0.4151	6.8037	4.7138
MELAKA	0.4569	12.9636	8.0051	0.0879	6.1645	4.3270
NEGERI SEMBILAN	0.7379	14.6435	12.1866	0.0891	12.2674	9.4876
PAHANG	0.0010	4.2375	3.3988	0.1300	4.6788	3.5125
PERAK	0.3109	11.3751	10.1358	0.3537	2.5815	2.1961
PULAU PINANG	0.3999	3.4504	2.7882	0.4159	3.7267	3.0677
SABAH	0.8016	19.0210	17.3570	0.0529	5.6880	3.7808
SARAWAK	0.3601	12.1759	9.9570	0.2862	7.8106	4.8680
SELANGOR	0.7699	17.7250	14.0839	0.4493	36.8844	31.6609
TERENGGANU	0.2581	15.9367	11.9460	0.4638	4.2410	3.6427
W.P. KUALA LUMPUR	0.4331	12.1759	9.9570	0.2886	23.9335	21.6772
W.P. LABUAN	0.9572	17.0286	14.9023	0.7878	1.8696	1.6406
W.P. PUTRAJAYA	0.2901	11.1573	8.8531	0.2050	3.4298	2.9108

From the prediction result, the lags of 10 days were also recorded in table 4.2. As mentioned before, this project studied the relationship between the number of death in one day with the number of cases from yesterday's up to 10 days. This lagging days are called as lags. The main reason is to observe the actual day of death once the patients are infected by COVID-19 disease. Figures shows the lags of 10 days for W.P. Labuan and Selangor with different trend of lags. However, the trend of the lags differs for each state depends on the errors and the size of the training and testing data for each of the state.

Based on observation in figure 4.6, the range of error between the two states are large. The trend of lags for W.P Labuan almost cannot be seen as the value of the errors are far smaller compared to Selangor. From Figure 4.6, the lags of day 2 has the lowest value of error which is 1.9762. This indicates that most patients in W.P. Labuan who were infected by COVID-19 disease are most likely to die on the second day after the infections.

On the other hand, the lags of day 6 has the lowest value of error which is 39.1678. This specifies that most patients in Selangor who were infected by COVID-19 disease are most likely to die on the sixth day after the infections. The evaluation of the lags of 10 days for each of the state were shown in table 4.2.

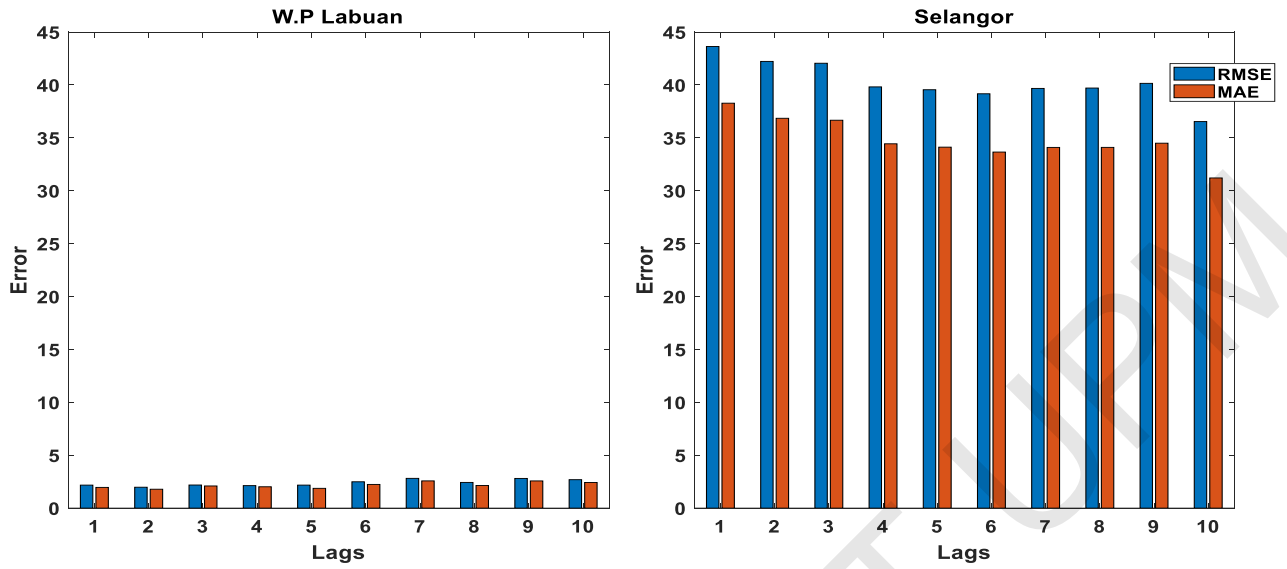


Figure 4.6 Trend of Lag of 10 days for W.P Labuan and Selangor

Table 4.2 Evaluation of Lag of 10 days for each state

State		JOHOR	KEDAH	KELANTAN	MELAKA	NEGERI SEMBILAN
Lag 1	cc	0.5891	0.2871	0.4180	0.4898	0.1750
	RMSE	17.4160	5.0143	6.9077	6.2507	9.6722
	MAE	15.3185	4.3966	4.7031	4.3956	7.7621
Lag 2	cc	0.5817	0.2561	0.4190	0.2794	0.1727
	RMSE	17.0803	5.0280	6.8384	5.7491	10.3078
	MAE	15.0052	4.4240	4.7040	4.1743	8.2131
Lag 3	cc	0.5804	0.3518	0.4089	0.4167	0.2082
	RMSE	17.0930	5.3207	6.8000	6.6639	10.0620
	MAE	15.0210	4.6906	4.7338	4.6270	8.0897
Lag 4	cc	0.5590	0.3686	0.2145	0.5047	0.1723
	RMSE	16.4641	5.1580	8.0182	5.5784	9.4936
	MAE	14.4315	4.5516	5.4614	3.8466	7.9126
Lag 5	cc	0.5619	0.3696	0.3905	0.3560	0.0356
	RMSE	16.4715	5.4309	6.7712	11.8872	10.7118
	MAE	14.4470	4.7894	4.8161	9.0580	9.0070
Lag 6	cc	0.5389	0.3607	0.3605	0.2504	0.0724
	RMSE	15.9191	5.9441	6.8715	5.8394	10.7871
	MAE	13.9330	5.1148	4.8972	4.1488	9.3417

Lag 7	cc	0.5214	0.2971	0.3260	0.3347	0.0955
	RMSE	15.6329	5.5685	6.9775	12.2124	13.0100
	MAE	13.6618	4.8791	4.9486	9.4474	10.9912
Lag 8	cc	0.5197	0.3773	0.2959	0.4171	0.0846
	RMSE	15.1485	6.4077	7.0093	5.4061	16.2861
	MAE	13.2324	5.6140	5.0671	3.9309	13.5012
Lag 9	cc	0.4510	0.4721	0.3160	0.3084	0.0855
	RMSE	14.6894	5.7888	7.0113	19.0306	20.1171
	MAE	12.8296	4.8682	4.9872	15.9681	17.8431
Lag 10	cc	0.4493	0.3732	0.2545	0.3051	0.2246
	RMSE	14.8469	7.5043	7.3625	31.0227	12.2037
	MAE	12.9688	6.7845	5.1373	26.8328	10.0089
State		PAHANG	PERAK	PULAU PINANG	SABAH	SARAWAK
Lag 1	cc	0.2304	0.1300	0.3999	0.0203	0.0332
	RMSE	4.5986	2.9709	5.0038	5.6564	12.8715
	MAE	3.3975	2.4525	4.2994	3.7807	8.0279
Lag 2	cc	0.1134	0.1286	0.3815	0.0393	0.2989
	RMSE	4.7184	4.1294	5.5604	5.6328	11.9540
	MAE	3.5358	3.5754	4.9953	3.7830	10.6187
Lag 3	cc	0.1696	0.1271	0.4306	0.0568	0.2922
	RMSE	4.5457	3.5082	2.7734	5.6056	11.8215
	MAE	3.3867	2.9135	2.2165	3.7887	10.4941
Lag 4	cc	0.1802	0.2310	0.3410	0.2060	0.3359
	RMSE	4.4590	4.7658	4.1865	5.5444	7.7519
	MAE	3.3460	4.1376	3.4337	3.7696	5.0163
Lag 5	cc	0.0811	0.2587	0.3716	0.2253	0.3008
	RMSE	4.4830	3.6905	3.9154	5.5354	11.9011
	MAE	3.5202	3.0565	3.1677	3.8039	10.5709
Lag 6	cc	0.0753	0.2075	0.3725	0.1866	0.2908
	RMSE	4.4604	5.1360	5.6323	5.5335	7.9184
	MAE	3.4488	4.5763	4.9293	3.8294	5.5055
Lag 7	cc	0.2675	0.0307	0.3422	0.1395	0.2065
	RMSE	4.1891	2.7691	5.8600	5.5528	7.9807
	MAE	3.3324	2.0648	5.1003	3.8411	5.3678
Lag 8	cc	0.0035	0.2137	0.3846	0.2189	0.2175
	RMSE	5.0136	6.2949	4.7172	5.5293	7.0407
	MAE	4.0503	5.5397	3.8631	3.8618	5.2070

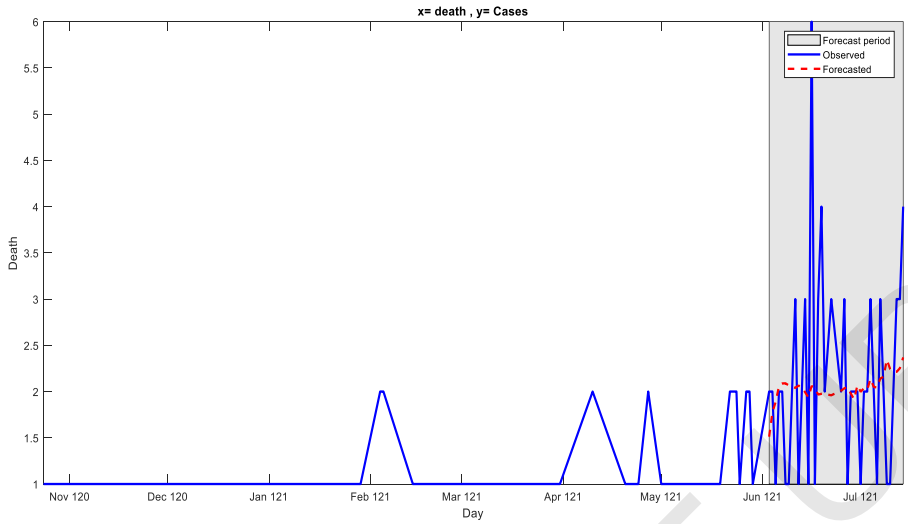
Lag 9	cc	0.2745	0.0653	0.3484	0.3142	0.1475
	RMSE	4.1452	2.8470	4.4215	5.5118	8.0832
	MAE	3.3091	2.1131	3.7255	3.8450	5.5184
Lag 10	cc	0.2804	0.1263	0.3549	0.3294	0.2890
	RMSE	4.1005	9.5461	3.9367	5.4980	7.9576
	MAE	3.4623	8.5629	3.2028	3.8810	5.4908
State		SELANGOR	TERENGGANU	W.P. KUALA LUMPUR	W.P. LABUAN	W.P. PUTRAJAYA
Lag 1	cc	0.4289	0.3964	0.2565	0.5654	0.7584
	RMSE	43.6387	5.1433	20.1437	2.1755	2.8640
	MAE	38.2782	4.6425	17.7375	1.9592	1.8513
Lag 2	cc	0.4266	0.3681	0.1415	0.6257	0.8670
	RMSE	42.2275	2.8714	16.3285	1.9762	2.2883
	MAE	36.8513	2.0237	12.4696	1.7871	1.7825
Lag 3	cc	0.4182	0.1488	0.2000	0.0112	0.0272
	RMSE	42.0549	6.6409	20.2457	2.1848	3.7616
	MAE	36.6684	6.1487	17.8993	2.0949	3.2788
Lag 4	cc	0.4108	0.5305	0.2377	0.5769	0.9463
	RMSE	39.8184	10.0164	22.0146	2.1266	1.1598
	MAE	34.4395	9.3721	19.7225	2.0185	0.9241
Lag 5	cc	0.4013	0.3857	0.1868	0.8426	0.8215
	RMSE	39.5503	8.7604	16.6285	2.1760	2.3787
	MAE	34.1233	8.1886	14.2441	1.8709	2.0366
Lag 6	cc	0.3947	0.3031	0.1489	0.7810	0.8746
	RMSE	39.1678	9.1090	19.2102	2.4901	2.7995
	MAE	33.6594	8.5038	16.7609	2.2398	1.9423
Lag 7	cc	0.3868	0.2798	0.2596	0.5899	0.4666
	RMSE	39.6748	9.2469	22.4118	2.8141	4.2761
	MAE	34.0984	8.6310	20.1690	2.5745	2.7345
Lag 8	cc	0.3793	0.5215	0.2731	0.8803	0.1927
	RMSE	39.7099	2.8655	24.4873	2.4299	8.5514
	MAE	34.1010	2.3403	22.3599	2.1428	8.1943
Lag 9	cc	0.3776	0.5345	0.0251	0.7219	0.8070
	RMSE	40.1534	2.9262	14.6203	2.8073	6.2829
	MAE	34.4976	2.3429	11.1526	2.5711	5.1867
Lag 10	cc	0.3779	0.5945	0.2623	0.4245	0.7756
	RMSE	36.5404	2.9982	27.3807	2.6885	3.0157
	MAE	31.2102	2.4086	25.1789	2.4263	2.1994

4.4 Prediction of probability of death of patients' age

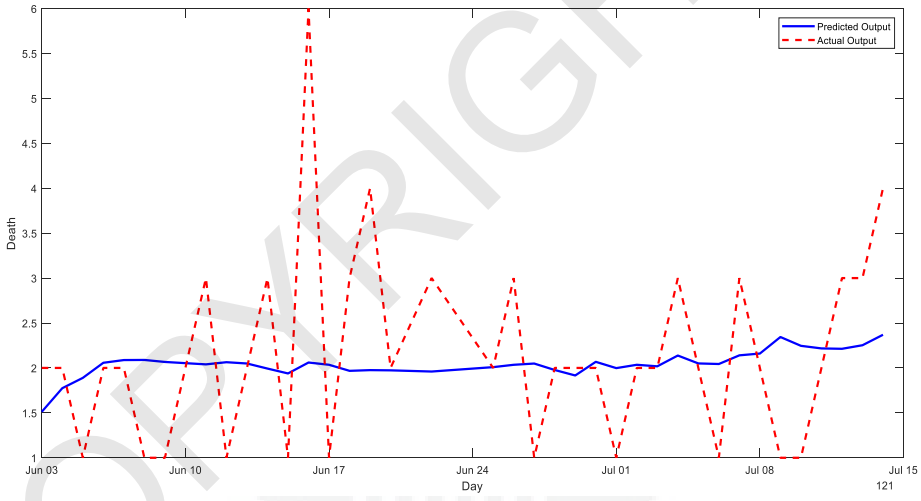
In this section, the result of prediction and the evaluation of the prediction of death for each of the age class were recorded and discussed. As mentioned in subchapter 4.2, the best-predicted result is when the value of cc is close to 1 with a lower value of RMSE and MAE.

From observation, the result of the prediction of death due to COVID-19 diseases for age 25-29 has the best prediction result with the lowest value of RMSE which is 1.0718 compared to the other class age. On the other hand, class age of 60-64 has been recorded with the highest value of RMSE which is 11.4184. The difference value of the error however depends on the size of the training and testing data. The size of the training and testing data for the class age 25-29 is 32 and 38, while the training and testing data for class age 60-64 which has the largest size of data set are 141 and 43 respectively.

The prediction of the probability of death due to COVID-19 disease for each class age are shown in the figures below. As shown in each figure 4.7 and 4.8, the blue line in graph (a) indicates the observed data, which is the testing data, while the red dotted line indicates the forecast data. The darker or gray area is basically specify the forecast period. For this case, the data from 1st June to 13th July are being forecasted using the testing data. On the other hand, for both figure (b), the blue line in graph indicates the predicted output and the red dotted line indicates the actual output.

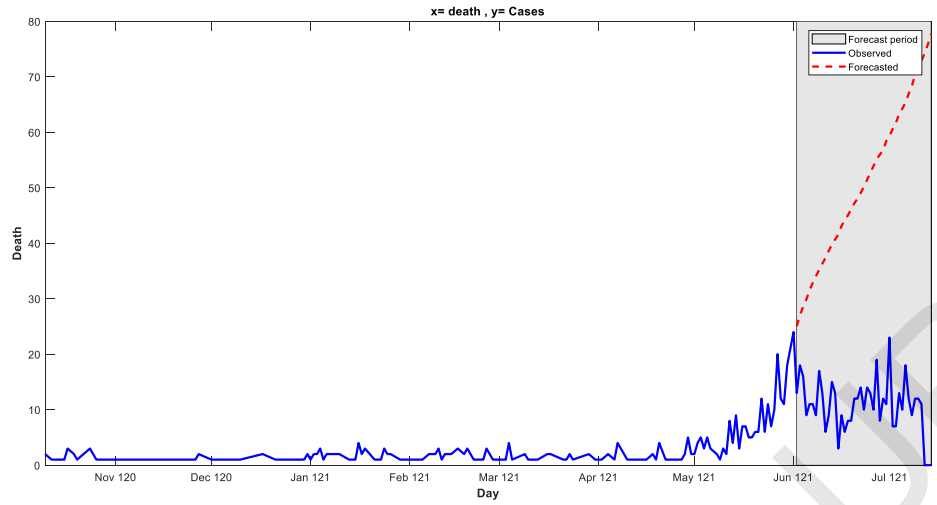


(a)

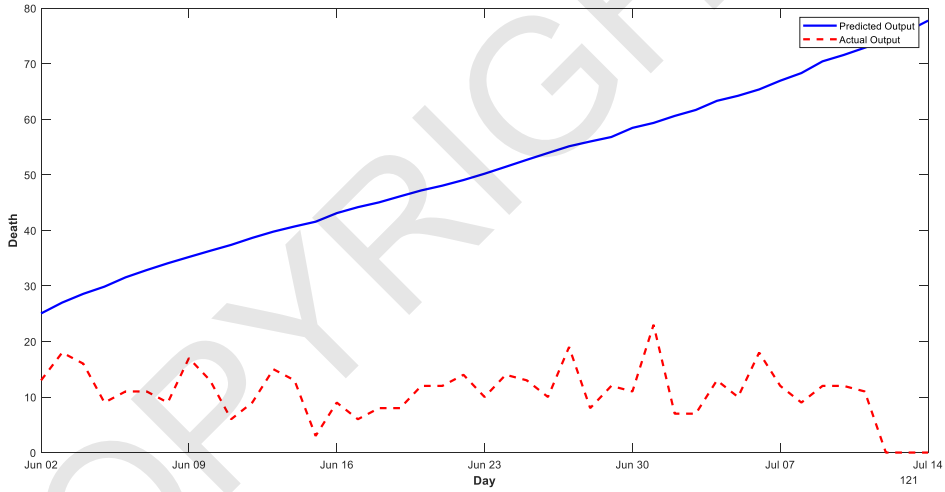


(b)

Figure 4.7 Prediction of death due to COVID-19 diseases for age 25-29



(a)



(b)

Figure 4.8 Prediction of death due to COVID-19 diseases for age 60-64

From the prediction result, the lagging data of 10 days were observed and recorded in table 4.3. Figures show the lags of 10 days for all the class age. From the observation each of the class age project different trend of lags. The trend of the lags differs for each class age depends on the errors. The value of RMSE and MAE differ based on the size of the training and testing data for each of the class age.

Based on figure 4.9, the lags for each of the class age have different trends. The error measured also recorded a wide different in range. As observed, the error for class age 25-29 are smaller compared to class age 60-64. The trend of lags for the class age 60-64 and 65-69 can be observed clearly compared to other class age, the trend of the lags can hardly be seen. This showed that the different range of the error between the class age are wide.

Based on the figure, each class age has different trends of lags. For example, class age 35-39 has recorded the lags of day 9 has the lowest value of RMSE and MAE which indicates that most patients in between the age 35-39 who were infected by COVID-19 disease are most likely to die on the ninth day after the infections. On the other hand, class age 60-64 has recorded the lags of day 1 has the lowest value of RMSE and MAE which indicates that most patients in between the age 60-64 who were infected by COVID-19 disease are most likely to die on the first day after the infections.

However, the errors and the trend of lags are affected by the size of the training and testing data. For instance, the class age 60-64 has the largest size of training and testing data which results in higher value of error compared to the other class age. The lags of each of the class age differ depends on the size of the training and testing data. As mentioned before, each of the class age has different trends of lags depends on the data size.

Overall, patients in their 60's and above tends to die in the first and second day after the infections. The evaluation of error and cc of the lags of 10 days for each class age were recorded in table 4.3.

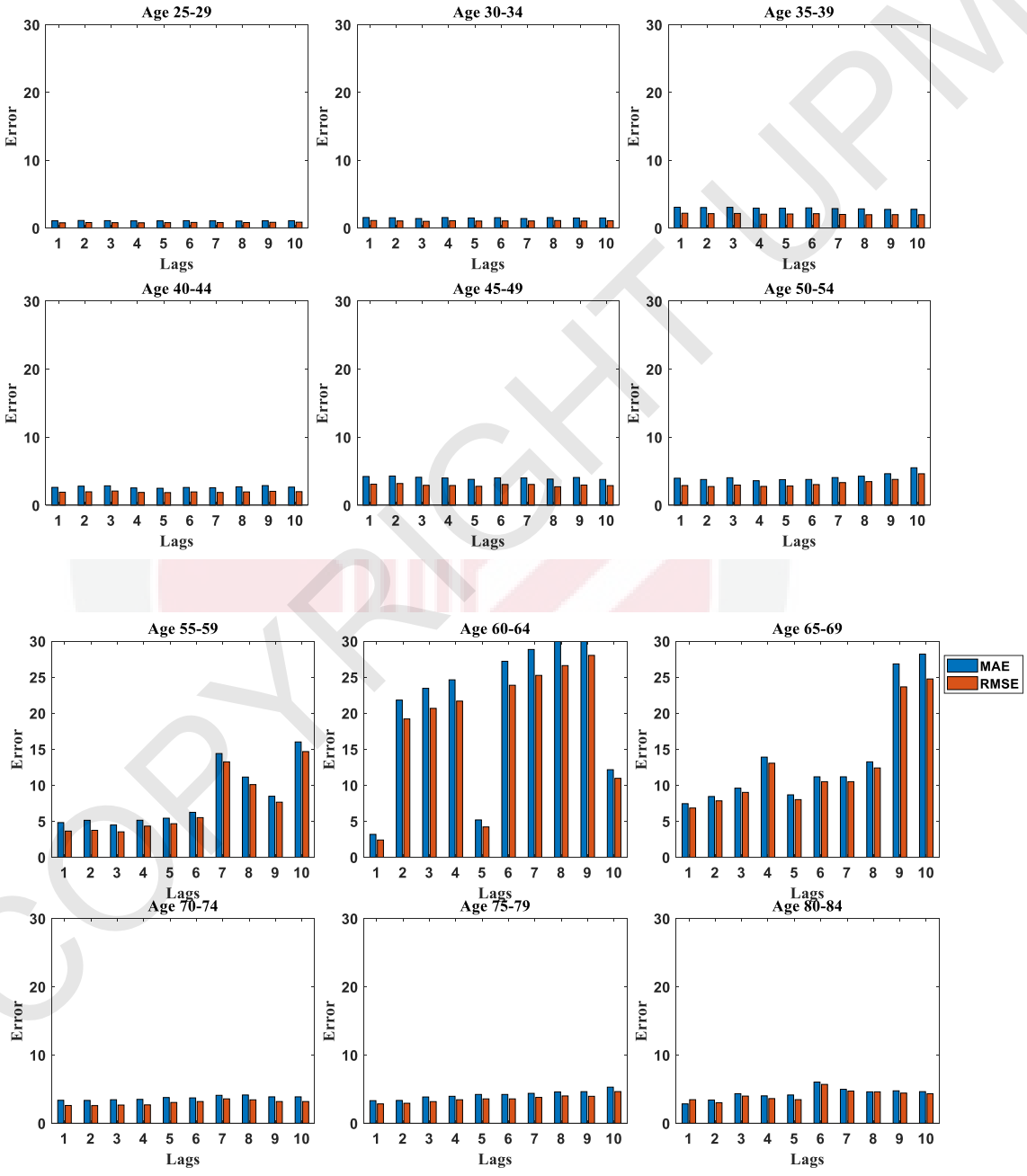


Figure 4.9 Trend of Lag of 10 days for each age class

Table 4.3 Evaluation of lags of 10 days for each of the class age

Age	Lag 1			Lag 2			Lag 3		
	cc	RMSE	MAE	cc	RMSE	MAE	cc	RMSE	MAE
25-29	0.105	1.065	0.769	0.179	1.109	0.807	0.002	1.085	0.788
30-34	0.044	1.547	1.107	0.080	1.510	1.044	0.263	1.401	0.988
35-39	0.028	3.064	2.178	0.043	3.020	2.118	0.036	3.047	2.145
40-44	0.100	2.626	1.916	0.008	2.803	1.974	0.086	2.853	2.087
45-49	0.312	4.226	3.090	0.064	4.290	3.191	0.098	4.108	2.928
50-54	0.182	3.970	2.901	0.243	3.782	2.752	0.172	4.045	2.961
55-59	0.280	4.810	3.630	0.230	5.130	3.740	0.476	4.476	3.527
60-64	0.678	3.177	2.395	0.151	21.815	19.197	0.155	23.446	20.667
65-69	0.020	7.438	6.833	0.118	8.432	7.830	0.185	9.592	9.003
70-74	0.097	3.364	2.602	-0.101	3.345	2.589	0.047	3.448	2.670
75-79	-0.020	3.310	2.839	0.111	3.342	2.931	0.164	3.850	3.171
80-84	-0.127	2.839	3.452	0.059	3.402	3.001	-0.075	4.330	3.981
Age	Lag 4			Lag 5			Lag 6		
	cc	RMSE	MAE	cc	RMSE	MAE	cc	RMSE	MAE
25-29	0.157	1.054	0.773	0.109	1.062	0.792	0.031	1.083	0.823
30-34	0.177	1.544	1.073	0.115	1.482	1.029	0.117	1.530	1.048
35-39	0.259	2.935	2.042	0.189	2.921	2.068	0.125	2.972	2.112
40-44	0.232	2.543	1.882	0.279	2.494	1.861	0.040	2.612	1.957
45-49	0.123	4.002	2.901	0.218	3.804	2.782	0.088	4.022	3.049
50-54	0.264	3.602	2.770	0.282	3.764	2.823	0.203	3.787	3.038
55-59	0.219	5.137	4.319	0.068	5.421	4.656	0.069	6.229	5.501
60-64	0.153	24.630	21.675	0.084	5.180	4.209	0.159	27.201	23.874
65-69	0.194	13.897	13.053	-0.018	8.665	8.000	0.278	11.166	10.492
70-74	0.059	3.496	2.696	-0.112	3.767	3.047	0.000	3.717	3.182
75-79	-0.068	3.950	3.441	0.072	4.220	3.570	-0.198	4.381	3.788
80-84	-0.126	4.013	3.614	0.013	4.141	3.456	0.111	6.033	5.690
Age	Lag 7			Lag 8			Lag 9		
	cc	RMSE	MAE	cc	RMSE	MAE	cc	RMSE	MAE
25-29	0.124	1.059	0.803	0.246	1.033	0.807	0.199	1.064	0.839
30-34	0.272	1.400	1.042	0.440	1.538	1.102	0.009	1.479	1.034
35-39	0.075	2.867	2.003	0.021	2.827	1.961	0.017	2.745	1.973
40-44	0.114	2.563	1.881	0.083	2.699	1.957	0.084	2.887	2.046
45-49	0.056	4.008	3.055	0.066	3.838	2.716	0.202	4.078	2.961

50-54	0.195	4.066	3.328	0.190	4.270	3.471	0.166	4.633	3.797
55-59	0.020	14.396	13.241	0.193	11.132	10.086	0.235	8.469	7.647
60-64	0.156	28.842	25.245	0.153	30.461	26.607	0.153	32.215	28.022
65-69	0.278	11.166	10.492	0.288	13.244	12.395	0.042	26.833	23.653
70-74	-0.142	4.091	3.572	-0.170	4.140	3.432	-0.088	3.869	3.182
75-79	-0.202	4.591	4.001	-0.345	4.629	3.950	-0.309	4.983	4.328
80-84	0.244	4.973	4.716	0.077	4.603	4.603	0.113	4.752	4.419

Age	Lag 10		
	cc	RMSE	MAE
25-29	0.214	1.076	0.857
30-34	0.077	1.479	1.069
35-39	0.125	2.761	1.963
40-44	0.082	2.662	1.994
45-49	0.133	3.794	2.888
50-54	0.153	5.506	4.626
55-59	0.043	16.005	14.668
60-64	0.386	12.153	10.955
65-69	-0.039	28.194	24.747
70-74	-0.088	3.869	3.182
75-79	-0.307	5.287	4.639
80-84	0.242	4.631	4.325

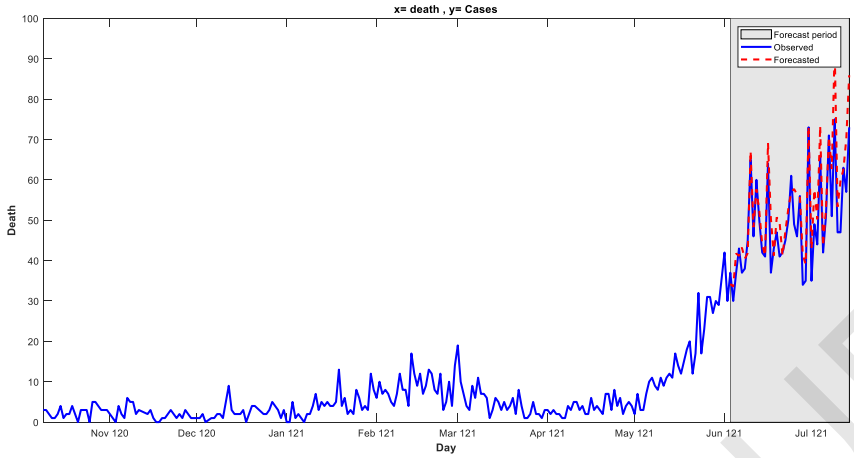
4.5 Prediction of probability of death based on patients' medical history

As mentioned in the previous chapter, there are 70 various types of diseases recorded in the data of patients' medical history. The five diseases with highest number of patients who died from COVID-19 are Hypertensive heart disease (28.5%) or also known as high blood pressure disease, followed by Diabetes (22.2%), Kidney disease (6.3%), Heart disease (8.3%), and Dyslipidemia (8.7%). The prediction result for each of the five mentioned diseases are shown in figures below.

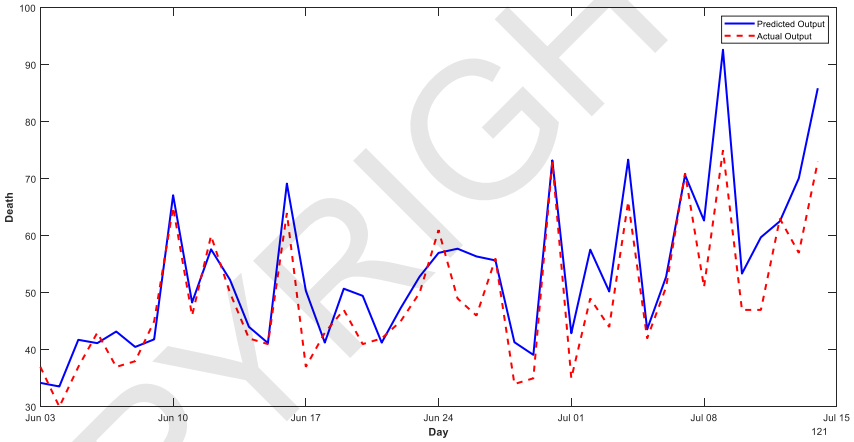
As shown in each figure, the blue line in graph (a) indicates the observed data, which is the testing data, while the red dotted line indicates the forecast data. The darker or gray area is basically specify the forecast period. For this case, the data from 1st June to 13th July are being forecasted using the testing data. On the other hand, for both figure (b), the blue line in graph indicates the predicted output and the red dotted line indicates the actual output.

Each of the figures shown projects different trends and results. From observation in Figure 4.10, the prediction of death due to COVID-19 disease for Dyslipidemia gave the lowest difference in predicted and actual value which is, 1.6462 compared to the other diseases. Dyslipidemia has recorded the lowest value of RMSE compared to the other diseases, which is 3.0629 with the value of cc, 0.8025.

Based on Table 4.4, Dyslipidemia has recorded the lowest value of RMSE and MAE, which are 3.0629 and 2.207 with the value of cc, 0.8025. On the other hand, Kidney disease has recorded the highest value of RMSE, which is 27.7412 with the lowest value of cc, 0.1312. From all of the diseases, the prediction of death due to COVID-19 of each of the mentioned diseases, Dyslipidemia has recorded the best result compared to the other diseases.

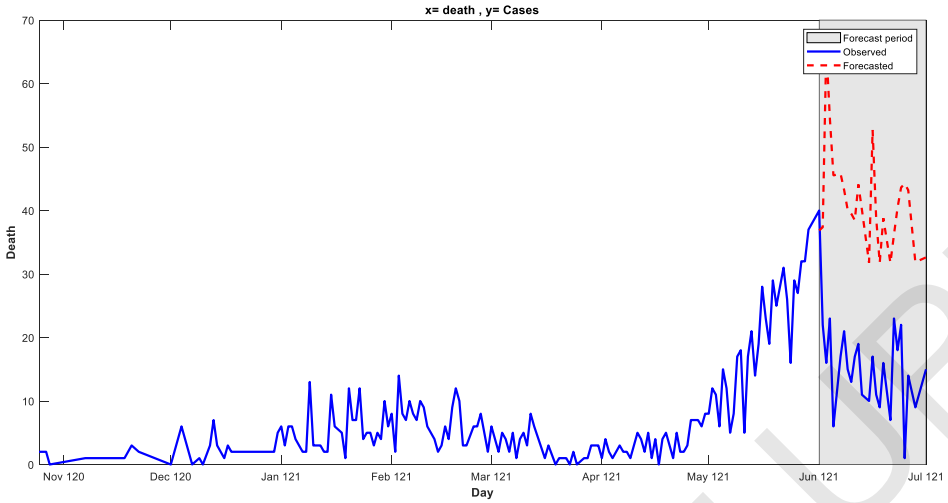


(a)

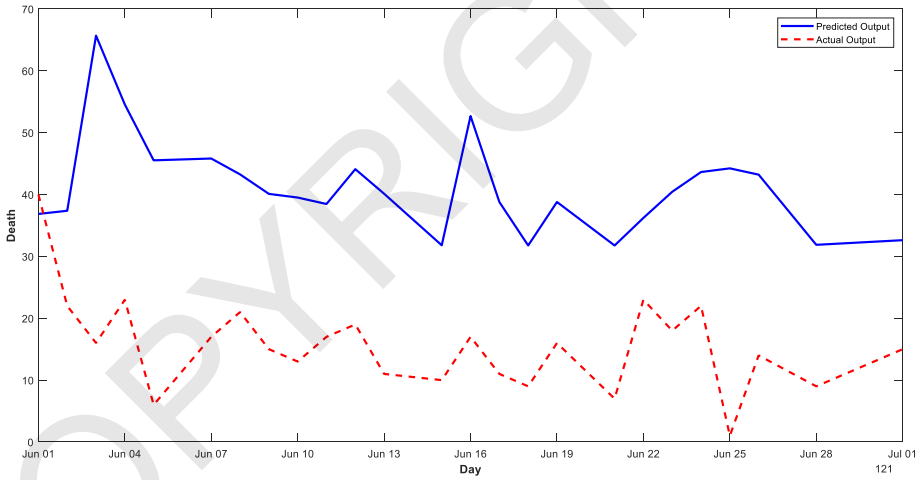


(b)

Figure 4.10 Prediction of death due to COVID-19 disease for Dyslipidemia



(a)



(b)

Figure 4.11 Prediction of death due to COVID-19 disease for Kidney disease

Table 4.4 Evaluation of prediction of death due COVID-19 for patients' background disease

Diseases	Evaluation		
	cc	RMSE	MAE
Hypertensive heart disease	0.9201	7.5595	5.9109
Diabetes	0.9250	5.5176	4.3044
Heart disease	0.6719	3.9286	3.3570
Dyslipidemia	0.8025	3.0629	2.2070
Kidney disease	0.1312	27.3844	25.7412

CHAPTER 5

CONCLUSION

5.1 Summary of the thesis

In conclusion, the main objective of the project is to determine which parameters give the best prediction result to forecast the probability of death due to COVID-19 disease. The prediction of probability of death due to COVID-19 disease for each state in Malaysia, patients' age and medical history using the time series data were performed using ARIMAX model. This work focuses more on studying the trend and relationship between the new confirmed cases and death cases for each of the mentioned factors.

Based on the statistical analysis, the trend of the new confirmed cases and death cases were increasing up until July 2020. The trend of the new confirmed cases can be seen to be correspond with data of death cases with certain time lag. The time lagging for each state and patients' age however depends on the size of the training and testing data. Overall, based on the cc value, the data of death cases correlates more with the data of death in each state, patients' age and medical history.

The data of death for each state, patient's age and medical history were divided into two parts, training and testing for prediction. Using the ARIMAX prediction model, the data of new confirmed cases gave lower value of error and higher value of Cross Correlation (cc). The value of errors and cc of prediction for each state, patients' age and medical history then were compared and the data of the medical history by far gave the best prediction result.

5.2 Future work

This work predicts the probability of death of COVID-19 for each state, patient's age and medical history using ARIMAX time series model. In future, further research could be based on different machine learning algorithms that utilizes the most latest and advanced algorithm for prediction. More time series model can be used to predict the trend of death of COVID-19 and compare which model would give the best predicted result. Also, another parameters such as the number of patients' close contact, patients' in ICU, vaccination rates and recovered cases should be taken into account for further study to understand more about the trend of death of COVID-19 in Malaysia.

REFERENCES

Moody, J. (2021, December 11). *What does RMSE really mean? - Towards Data Science*. Medium. https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e

M, E. |. (2018, April 7). *Mean Absolute Error ~ MAE [Machine Learning(ML)] - E | M*. Medium. https://medium.com/@ewuramaminka/mean-absolute-error-mae-machine-learning-ml-b9b4afc63077

Kadel, S., & Kovats, S. (2018). Sex Hormones Regulate Innate Immune Cells and Promote Sex Differences in Respiratory Virus Infection. *Frontiers in Immunology*, *9*. https://doi.org/10.3389/fimmu.2018.01653

Techopedia. (2021, July 11). *Data Preprocessing*. Techopedia.Com. https://www.techopedia.com/definition/14650/data-preprocessing

Takahashi T, Ellingson MK, Wong P, Israelow B, Lucas C, Klein J, et al. Sex differences in immune responses that underlie COVID-19 disease outcomes. Nature. 2020

Purwandari, T., Zahroh, S., Hidayat, Y., Sukonob, S., Mamat, M & Saputra, J. (2022). Forecasting model of COVID-19 pandemic in Malaysia: An application of time series approach using neural network.*Decision Science Letters* , 11(1), 35-42.

Shaharudin, S. M., Ismail, S., Hassan, N. A., Tan, M. L., & Sulaiman, N. A. F. (2021). Short-Term Forecasting of Daily Confirmed COVID-19 Cases in Malaysia Using RF-SSA Model. *Frontiers in Public Health*, *9*. https://doi.org/10.3389/fpubh.2021.604093

da Silva, T. T., Francisquini, R., & Nascimento, M. C. (2021). Meteorological and human mobility data on predicting COVID-19 cases by a novel hybrid decomposition method with anomaly

detection analysis: A case study in the capitals of Brazil. *Expert Systems with Applications*, *182*, 115190. https://doi.org/10.1016/j.eswa.2021.115190

Li, Y., Horowitz, M. A., Liu, J., Chew, A., Lan, H., Liu, Q., Sha, D., & Yang, C. (2020). Individual-Level Fatality Prediction of COVID-19 Patients Using AI Methods. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.587937

Jinming Cao, Xia Jiang, Bin Zhao, et al. Mathematical Modeling and Epidemic Prediction of COVID-19 and its Significance to Epidemic Prevention and Control Measures. J BioMed Res Innov. 2020;1(1):103.

Perone, G. (2020). ARIMA forecasting of COVID-19 incidence in Italy, Russia, and the USA. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3612402

Yang, Q., Wang, J., Ma, H., & Wang, X. (2020). Research on COVID-19 based on ARIMA modelΔ—Taking Hubei, China as an example to see the epidemic in Italy. *Journal of Infection and Public Health*, *13*(10), 1415–1418. https://doi.org/10.1016/j.jiph.2020.06.019

Wu, Y. C., Chen, C. S., & Chan, Y. J. (2020). The outbreak of COVID-19: An overview. *Journal of the Chinese Medical Association*, *83*(3), 217–220. https://doi.org/10.1097/jcma.0000000000000270

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 1994.

Wold, H. *A Study in the Analysis of Stationary Time Series*. Uppsala, Sweden: Almqvist & Wiksell, 1938.

Elengoe, A. (2020b). COVID-19 Outbreak in Malaysia. *Osong Public Health and Research Perspectives*, *11*(3), 93–100. https://doi.org/10.24171/j.phrp.2020.11.3.08

APPENDICES

MATLAB ARIMAX coding :

```
T = size(Data_ForecastDisease171,1) % Total sample size
GNPR = Data_ForecastDisease171.Var2;
varnames = [ "Var4" ]; %boleh tukar ikut varibale anda mahukan
X = Data_ForecastDisease171{:,varnames(1:end)};
idxpresample = 1;
idxestimate = 2:160; %traindata
idxforecast =161:T; %testingdata



dates = datetime(m,'ConvertFrom',"datenum",...
    'Format',"yyyy-MM-dd");
Mdl = arima(1,0,2);

y0est = GNPR(idxpresample); % Presample response data for estimation
yest = GNPR(idxestimate);   % Response data for estimation
XEst = X(idxestimate,:);     % Estimation sample exogenous data

Mdl = estimate(Mdl,yest,'Y0',y0est,'X',XEst,'Display','off');

y0f = yest((end - 2):end); % Presample response data for forecasting
X0f = XEst((end - 1):end,:); % Presample exogenous data for forecasting
      XF = X(idxforecast,:);     % Forecast period exogenous data for
model regression component
yf = forecast(Mdl,25,y0f,'X0',X0f,'XF',XF); %30=dataTest
yrs = dates(1:end); % k=1:T yrs=k(1:end)

ytest = GNPR(idxforecast,:);
cc=corrcoef([ytest,yf]);
cctest=cc(2);
RMSEtest= sqrt(mean(abs(ytest - yf).^2));
MAEtest= mean(abs(ytest - yf));

testing=[ytest,yf]
testingperformance=[cctest,RMSEtest,MAEtest]
MAPEtr=mean((((abs(ytest - yf))./ytest))*100);
```

```
figure;
plot(yrs(end-24:end),yf,"b","LineWidth",2)
xlabel("Day",'FontWeight','bold')
ylabel("Death",'FontWeight','bold')



hold on
plot(yrs(end-24:end),ytest,"r--","LineWidth",2)
'FontWeight','bold'
legend([ "Predicted Output" "Actual Output"])
hold off

figure;
plot(yrs,GNPR(1:end),"b","LineWidth",2);
hold on
plot(yrs(end-24:end),yf,"r--","LineWidth",2); % 30-1
h = gca;
px = yrs([end - 24 end end end - 24]);% 30-1
py = h.YLim([1 1 2 2]);
hp = patch(px,py,[0.9 0.9 0.9]);
uistack(hp,"bottom");
axis tight
title("x= death , y= Cases");
xlabel("Day",'FontWeight','bold')
ylabel("Death",'FontWeight','bold')
legend(["Forecast period" "Observed" "Forecasted"])
```

MATLAB Colourmap coding :

```
B=corr(d);

ii = ones(size(B));
idx = tril(ii);
B(~idx) = nan;
yourlabelnames=({'Death Cases','New Cases','Hypertensive heart
disease','Diabetes','Heart disease','Kidney disease','Dyslipidemia'});
%yourlabelnames=CovidData2.Properties.VariableNames;
%set(gca, 'XTickLabel', yourlabelnames); % set x-axis labels
%set(gca, 'YTickLabel', yourlabelnames); % set y-axis labels
heatmap(yourlabelnames,yourlabelnames,B, 'MissingDataColor', 'w');%,
%'GridVisible', 'off', 'MissingDataLabel', " ")
```

Data Notations

Diseases :

Scientific Disease name	Notations
Addison's disease	1
Adrenal insufficiency	2
Alzheimer's	3
Anemia	4
Arrhytmia heart disease	5
Asthma	6
Autoimmune disorders	7
Back pain	8
Benign prostatic hyperplasia (BPH)	9
Blind (OKU)	10
Breast cancer	11
Bronchiectasis	12
Cancer	13
Cataract	14
Cervical cancer	15
Chronic hepatitis B	16
Chronic kidney disease	17
Chronic obstructive airway disease	18
Chronic obstructive pulmonary disease (COPD)	19
Chronic pulmonary disease	20
Dementia	21
Depressive disorder	22
Diabetes	23
Down syndrome	24
Dry cough	25
Dyslipidemia	26
Gallstone disease	27
Gastritis	28
Gout	29
Heart disease	30
Hepatocellular carcinoma (HCC)	31

Hydrocephalus	32
Hypertensive heart disease	33
Hyperthyroidism	34
Hypothyroidism	35
Immunodeficiency disorders	36
Kidney disease	37
Leukemia	38
Liver disease	39
Lung cancer	40
Lymphoma	41
Neurogenic bladder	42
Obesity	43
Osteoarthritis	44
Pancreatic cancer	45
Paralysis	46
Parkinson	47
Peripheral vascular disease	48
Psoriasis	49
Retroviral	50
Rheumatoid arthritis	51
Schizophrenia	52
Sinusitis	53
Skin cancer	54
Sleep apneia syndrome	55
Spondylodisctis	56
Stroke	57
Systemic lupus erythematosus (SLE)	58
Thyroid cancer	59
Thyroid disease	60
Tuberculosis (TB)	61
Thalassemia	62
Cerebral Palsy	63
Deep vein thrombosis	64
Myelodysplastic Syndrome	65
Glioblastoma Multiforme	66
Muscular dystrophy	67
Deaf	68
Osteoporosis	69
Hemophilia	70
Not related to any diseases	71
Not available	72

State:

State	Notation
Johor	1
Kedah	2
Kelantan	3
Melaka	4
Negeri Sembilan	5
Pahang	6
Pulau Pinang	7
Perak	8
Perlis	9
Selangor	10
Terengganu	11
Sabah	12
Sarawak	13
WP Kuala Lumpur	14
WP Labuan	15
WP Putrajaya	16

Sex:

Sex	Notation
Male (M)	1
Female (F)	2

VITAE