A unified view of epidemiological modelling and artificial intelligence-based approach to understanding the COVID-19 disease spread for supporting public policy decisions

by

ALI BOUKRICH

A project report submitted in conformity with the requirements
for the degree of Master of Science in Information Technology

Department of Mathematical and Physical Sciences (Graduate Studies)
Faculty of Graduate Studies
Concordia University of Edmonton

A UNIFIED VIEW OF EPIDEMIOLOGICAL MODELLING AND ARTIFICIAL INTELLIGENCE-BASED APPROACH TO UNDERSTANDING THE COVID-19 DISEASE SPREAD FOR SUPPORTING PUBLIC POLICY DECISIONS

ALI BOUKRICH

Approved:

Supervisor: Baidya Nath Saha, Ph. D. Date

Committee Member Date

Dean of Graduate Studies: Alison Yacyshyn, Ph. D. Date

A unified view of epidemiological modelling and artificial intelligence-based approach to understanding the COVID-19 disease spread for supporting public policy decisions

ALI BOUKRICH
Master of Science in Information Technology
Department of Mathematical and Physical Sciences (Graduate Studies)
Concordia University of Edmonton
2022

Abstract

The Covid-19 epidemic has emerged as one of the most concerning global public health catastrophes of the twenty-first century, highlighting the critical need for robust forecasting approaches for disease identification, alleviation, and prevention, among other things. Forecasting is one of the most powerful statistical methods for detecting and evaluating trends and forecasting future consequences based on which timely and mitigating actions can be performed all over the world in numerous disciplines. Several statistical methodologies and machine learning techniques have been employed to this goal, depending on the study needed and the data available. Most of the predictions made in the past have been short-term and country-specific. In this paper, an assessment of the potential machine learning technique is suggested for forecasting Covid-19-related characteristics in the long run, both in Canada and globally. This recommended ML model seems to be well for forecasting data from the past and present. Three datasets were used in this analysis, from the Alberta Health Services, Statistics Canada, and Worldometers, respectively. Long-term data forecasts for both Alberta and Canada were detailed using these three datasets, and it was discovered that anticipated data was highly similar to real-time values. The experiment was also carried out for Canadian province predictions as well as country-level predictions around the world, and the results are presented in the Appendix [1].

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Coronavirus disease 2019 (COVID-19) has emerged as a major public health concern around the world. There have been 511,965,711 confirmed cases of COVID-19 reported to WHO as of 5:46pm CEST on 3 May 2022, with 6,240,619 deaths. Globally a total of 11,560,378,840 vaccine doses have been delivered as of May 3, 2022. In Canada, from 3 January 2020 to 5:46 pm CEST, 3 May 2022, there have been 3,753,470 confirmed cases of COVID-19 with 39,289 deaths, reported to WHO. As of 29 April 2022, a total of 81,841,579 vaccine doses have been administered. [2]. The COVID-19 pandemic has emerged at remarkable speed, and it is likely of a bat-origin that may have been transmitted to humans. The virus was likely already capable of human-to-human transmission but evolved more efficient transmissibility in late 2019. The human-to-human transmission was officially recognized by the global public health community in mid-January 2020 [3]. Intensive public health measures such as case detection, contact tracing and quarantine, as well as social distance, were initiated shortly after. In Canada, the four largest provinces (British Columbia, Alberta, Ontario, and Quebec) have recorded the bulk of cases and deaths, and physical separation (including school, college, and university closures, as well as "non-essential" company closures) was adopted starting in mid-March 2020, and consequent decreases in disease transmission are decreasing the outbreak [4]. However, provinces are free to take a separate decision about there response to the outbreak.

1.2 Problem Statement

In Canada and between January 2020 and May 2022, the world health organization (WHO) trend on Covid and the public health and social measures (PKSM) show

how consistent the trend is and policy decisions. We can see that the number of cases and health measures are related, but we can not know if the rise of cases forces decision makers to take action or if the health measures decrease the number of cases. The latest new coronavirus disease (COVID-19) outbreak, which is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is seeing a dramatic rise in infected individuals all over the world. The immunological response of the host to SARS-CoV-2 appears to be important in the illness etiology and clinical symptoms. In patients with severe COVID-19, SARS-CoV-2 not only triggers antiviral immune responses, but can also trigger uncontrolled inflammatory responses defined by high levels of pro-inflammatory cytokines, resulting in lymphopenia, lymphocyte dysfunction, and granulocyte and monocyte abnormalities. These immunological abnormalities caused by SARS-CoV-2 could lead to microbial infections, septic shock, and severe multiple organ failure. As a result, the processes underlying immunological abnormalities in COVID-19 patients must be understood to guide clinical care of the disease.[5].

1.3 Contribution of the thesis

To train and assess several non–time series machine learning models in predicting confirmed infection growth, we integrated the Alberta COVID-19 Government Response Tracker data set with Canada's daily reported COVID-19 infection case numbers. Our findings show that when the government did not take action to control the spread, the transmission rate Rt was high, and when the government did take action to limit the spread, the transmission rate Rt was low.
The research task entails:

- Gathering, cleaning, and analyzing data to extract meaningful insights and information. We first start with one look at the data used in this project. All data was collected from the Alberta-health service website, and figure 2 shows the cleaned data only for Canada using Pandas.

- The collected data is prepossessed to see the Covid-19 Disease Spread across Canada and around the world. We will have one idea about the position of Canada between other countries based on their GDP.

- Epidemiological Modelling for COVID-19 Disease Prediction and Machine Learning based COVID-19 Prediction for Canada is done using a machine learning technique. The methodologies utilized to evaluate the outbreak when it begins are crucial to intervening steps to eradicate such deadly diseases. The patterns that appear in such settings are usually non-linear, which pushes us to create a system that can record such non-linear dynamic changes. We can characterize the transmission of infectious diseases with the help of these non-linear systems [6].

- Bayesian Analysis for COVID-19 Prediction and Unifying the epidemiological and Artificial Intelligence based Modelling for COVID-19 disease prediction. These methods were performed on a portion of the data, the train data, to create a model that can be used to test the remaining data. The Bayesian optimization method improves forecasting performance by automatically selecting the appropriate hyperparameters for each model. On the other hand, long short-term memory (LSTM) is a deep learning artificial recurrent neural network (RNN) architecture. Unlike standard feedforward neural networks, LSTM contains feedback links. The vanilla neural networks (such as MLP) do not have the sequential processing power. However, there is an extension of feedforward neural networks for this purpose, called recurrent neural networks, where at each step, the input from the current time and the hidden state from the previous timestamp is used to make a prediction.

- Time Series Analysis for COVID-19 Disease Prediction. In this section, we are using the Prophet Forecasting Model and the ARIMA Forecasting Model to compare the output performance and accuracy using data sets containing confirmed cases from the Alberta website. Then we compare the forecasting model with the last 2 weeks of real data. Our results show that Prophet is better than ARIMA.

1.4 Organization of the thesis

The goal of this project is to combine ideas from the aforementioned articles and use machine learning to project them onto Alberta's population. We hope to make progress in the creation of critical computational tools for epidemiological research in general, and COVID-19 research in particular. The project will be divided into twelve sections. **Chapter 1** introduces the problem and the contribution of this work. **Chapter 2** gives a brief overview of the related works and the data used in this paper. **Chapter 3** describes how ML can be used to analyze Covid-19 in Canada and the world. **Chapter 4,5** describes different ML techniques for prediction and forecasting, including a general ML process flowchart. **Chapters 6, 7, 8, 9** describe the proposed symptoms-based prediction model for the classification of Covid-19 infection and the

ARIMA model for forecasting the future confirmed case count of Covid-19 in Alberta. **Chapter 10** is a Control Chart and Filtering for COVID-19 Disease Prediction. **Chapter 11** is a conclusion and future work.

Chapter 2

Literature Review

In this chapter, we present research efforts for Covid-19 disease spread prediction available in the literature. Prediction Models available in the literature are categorized into different classes which are discussed below.

2.1 Epidemiological Models

One of the Covid-19 epidemiological models was discussed by Mohammad and Masud and others in [7]. This study used a mathematical epidemic model (MEM), a statistical model, and recurrent neural network (RNN) versions to forecast the cumulative confirmed cases. We suggested a replicable approach for RNN variations that leveraged z-score outlier identification to address the stochastic character of RNN variants. We used Poisson likelihood fitting to quantify heterogeneity in susceptibility in the MEM, taking into account lockdowns and the dynamic dependency of the transmission and identification rates. The MEM provided extensive insights into the virus transmission and potential control tactics, while the experimental results revealed the superiority of RNN variants in forecasting accuracy.

2.2 Machine Learning-based Approach

The goal of this study is to use different machine learning techniques to predict COVID-19 severity at admission (LR). From January 26 to March 28, 2020, a retrospective design was used at JinYinTan Hospital. Fifty-eight demographic, clinical, and laboratory characteristics were chosen using the LassoCV method, Spearman's rank correlation, expert comments, and literature evaluation. To predict severe COVID-19, RF, SVM, and LR were used, and the models' performance was compared using the area under the curve (AUC) to see how they compare to each other. The top performance model also looked at the importance of features to determine severity. [8]

The usual methodologies are failing to correctly estimate the global consequences due to a lack of precise Covid-19 records and uncertainty. To address this problem, the study by [9] provides a meta-analysis based on Artificial Intelligence that predicts the global trend of the outbreak. Nave Bayes exhibited promising results with fewer Mean Absolute Error (MAE) and Mean Squared Error (MSE) than the other two machine learning techniques studied.

2.3 Deep Learning-based Models

Deep learning via LSTM models for COVID-19 infection forecasting in India is the first study that we are presenting. It identifies COVID-19 hotspots in Indian states, captures the initial (2020) and second (2021) waves of infections, and presents a two-month projection. Its model suggests that another wave of infections in October and November 2021 is unlikely; nonetheless, authorities must remain watchful due to new virus variations. The method's applicability in various countries and areas is motivated by the accuracy of the predictions [10].

The second study used deep learning-based models a novel approach based on combining deep learning models with statistical methods for COVID-19 time series forecasting by Abbasimehr and others. This research employs time series augmentation techniques to construct new time series that incorporate the original series' properties. The suggested strategy considerably increases the performance of long short-term memory and convolutional neural networks in terms of symmetric mean absolute percentage error and root mean square error measurements. The method employs three deep learning techniques in the context of COVID-19 time series forecasting [11].

2.4 Time series analysis

Hu, Nan and Nassar and others addressed The impact of the COVID-19 pandemic on pediatric health service use one year after the first pandemic outbreak in New South Wales Australia. This study compared the observed and predicted numbers of inpatient admissions and emergency department visits for chronic, acute infections, and injury conditions for each month during the COVID-19 period (January 2020-February 2021). It was based on data from two major pediatric hospitals in New South Wales (NSW) Australia. All of the analyses were done with autoregressive error models and stratified by patient age, gender, and socioeconomic position [12].

2.5 Bayesian Models

The Bayesian structural time series model (BSTS) is used in Xie, and Liming's work to investigate and predict total confirmed cases of COVID-19 infection in the United States from February 28, 2020, to April 6, 2020. Days, confirmed cases, daily, death cases daily, and fatality rates are among the factors considered. The author takes advantage of the flexibility of Local Linear Trend, Seasonality, and contemporaneous covariates of dynamic coefficients. The total number of confirmed cases of COVID-19 infection will continue to rise steadily, with the total number in the United States breaking over 600,000 shortly (in the subsequent months) Then, around mid-May 2020, you'll hit the pinnacle. In addition, the model predicts that the daily likelihood of variable Recovered cases is 0.07 [13].

Another Bayesian hierarchical spatial Model is presented by Chen, Jinjie and others. This study uses a Bayesian hierarchical model to investigate the impact of over-reporting and under-reporting at the state level in the United States. Misclassification correction necessitates the insertion of new parameters that are not directly identifiable by the observed data. The model incorporates spatial dependency as well as the influence of various factors on under-reporting and accurate incidence rates. It investigates the impact of over-reporting (false positives) in addition to under-reported (false negatives) false positives. Priors that are instructive are essential, and R algorithms that turn expert data into the proper prior distribution are discussed [14].

2.6 Spatio-temporal Analysis

Unemployment and population density were among the most influential variables with the highest relevance scores in terms of COVID-19 prevalence. Health-related variables such as diabetes prevalence and the number of hospital beds were also important predictors for mortality. The study by Kianfar, Nima and others used ten different variable importance analysis approaches to determine the relative importance of the explanatory variables. The outcomes of this study may provide general insights for public health policymakers who want to track illness spread and make better decisions [15].

From another perspective, we also analyze the case study on COVID-19 data discussed by Briz-Redón, Álvaro and others, which used a comparison of multiple neighbourhood matrix specifications for Spatio-temporal model fitting. This research compares and contrasts two situations. Using various neighbourhood matrices, modelling the weekly relative risk of COVID-19 over small areas in or near Valencia, Spain.

It generates neighbourhood matrices based on proximity, distance, covariate (mobility flows and sociodemographic characteristics), and hybrid matrices. It measures the goodness of fit, overall predictive quality, ability to detect high-risk Spatio-temporal units, ability to capture Spatio-temporal autocorrelation in the data, and goodness of smoothing for a collection of Spatio-temporal models based on each of the neighbourhood matrices. Matrixes based on proximity, some distance-based matrices, and those based on sociodemographic variables outperform matrices based on k-nearest neighbours and mobility flows, according to the findings [16].

2.7 Control Chart and Filtering

Finally, a study by Jahja, Maria and others, proposes a technique to estimate the daily number of new symptomatic COVID-19 infections at the county level in the United States. It concentrates on estimating infections in real-time (rather than retrospectively), which presents several difficulties. To address these issues, the authors create novel techniques for both the distribution estimation and deconvolution phases [17].

Chapter 3

Covid-19 Disease Spread across Canada and around the world.

3.1 Covid-19 data

COVID-19 datasets that are publically available are extremely difficult to come by due to privacy concerns, making research and development of AI-powered COVID-19 diagnosis tools problematic. To overcome this problem, we used open-source data from Statistics Canada, Alberta Health Services, CSSEGISandData from Github and Worldometer to conduct our research.

From the world data, we selected some countries that have almost the same GDP, HDI, HE, EI and PD. See figure 3.1

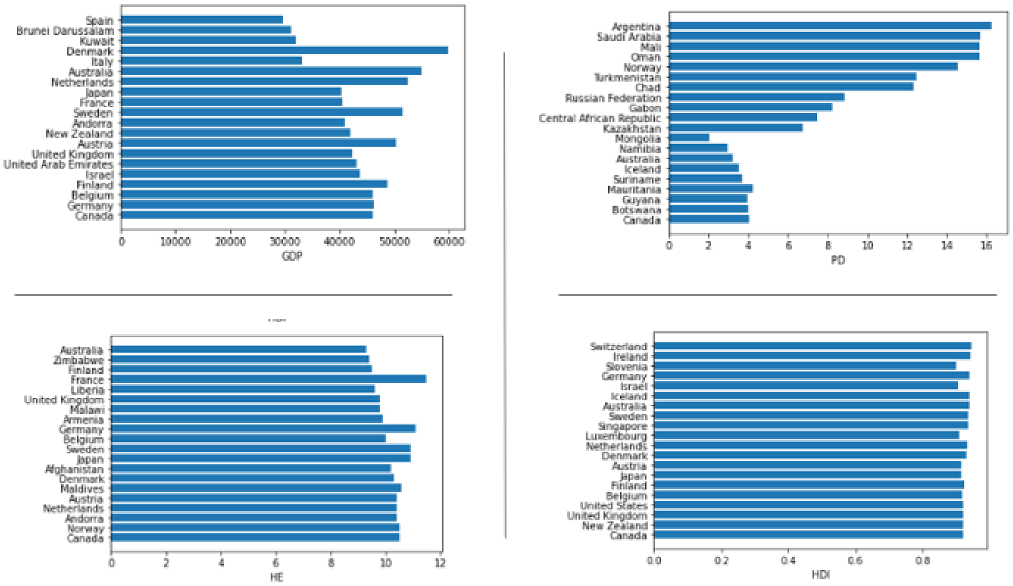
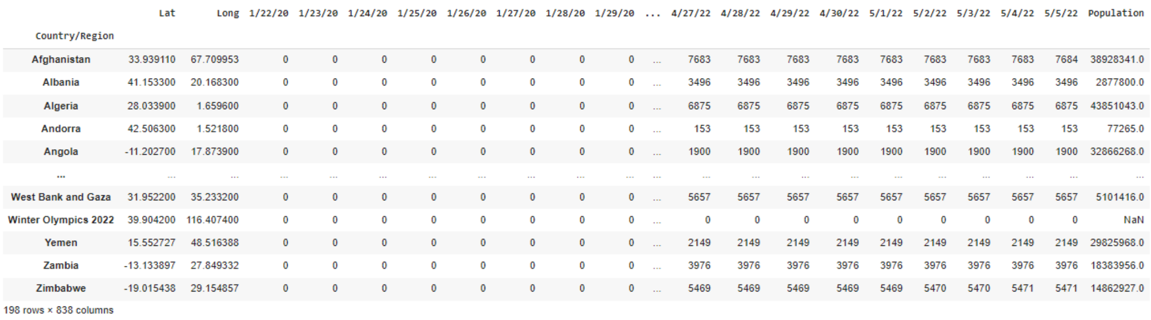


Figure 3.1: Selecting countries to compare to Canada

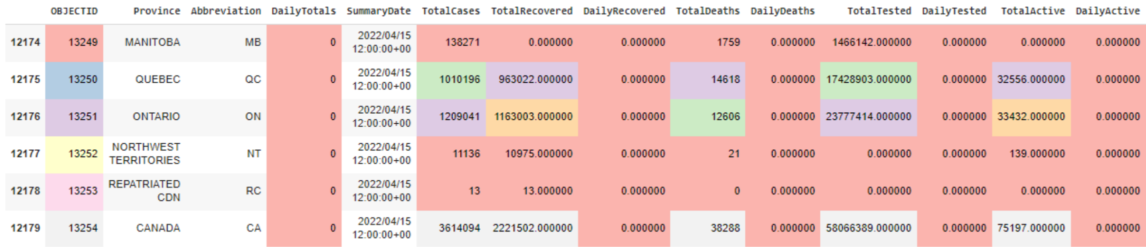
We combined the Alberta COVID-19 Government Response Tracker data set and Canada's daily reported COVID-19 infection case numbers to train and evaluate different non–time series machine learning models in predicting confirmed infection growth.



	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	...	4/27/22	4/28/22	4/29/22	4/30/22	5/1/22	5/2/22	5/3/22	5/4/22	5/5/22	Population
Country/Region																					
Afghanistan	33.939110	67.709953	0	0	0	0	0	0	0	0	...	7683	7683	7683	7683	7683	7683	7683	7683	7684	38928341.0
Albania	41.153300	20.168300	0	0	0	0	0	0	0	0	...	3496	3496	3496	3496	3496	3496	3496	3496	3496	2877800.0
Algeria	28.033900	1.659600	0	0	0	0	0	0	0	0	...	6875	6875	6875	6875	6875	6875	6875	6875	6875	43851043.0
Andorra	42.506300	1.521800	0	0	0	0	0	0	0	0	...	153	153	153	153	153	153	153	153	153	77265.0
Angola	-11.202700	17.873900	0	0	0	0	0	0	0	0	...	1900	1900	1900	1900	1900	1900	1900	1900	1900	32866268.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
West Bank and Gaza	31.952200	35.233200	0	0	0	0	0	0	0	0	...	5657	5657	5657	5657	5657	5657	5657	5657	5657	5101416.0
Winter Olympics 2022	39.904200	116.407400	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	NaN
Yemen	15.552727	48.516388	0	0	0	0	0	0	0	0	...	2149	2149	2149	2149	2149	2149	2149	2149	2149	29825968.0
Zambia	-13.133897	27.849332	0	0	0	0	0	0	0	0	...	3976	3976	3976	3976	3976	3976	3976	3976	3976	18383956.0
Zimbabwe	-19.015438	29.154857	0	0	0	0	0	0	0	0	...	5469	5469	5469	5469	5469	5470	5470	5471	5471	14862927.0

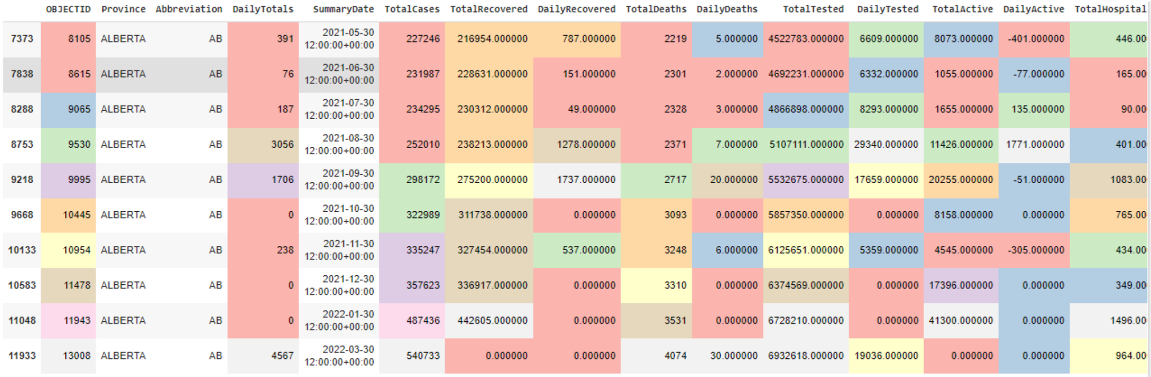
198 rows × 838 columns

Figure 3.2: World data.



	OBJECTID	Province	Abbreviation	DailyTotals	SummaryDate	TotalCases	TotalRecovered	DailyRecovered	TotalDeaths	DailyDeaths	TotalTested	DailyTested	TotalActive	DailyActive
12174	13249	MANITOBA	MB	0	2022/04/15 12:00:00+00	138271	0.000000	0.000000	1759	0.000000	1466142.000000	0.000000	0.000000	0.000000
12175	13250	QUEBEC	QC	0	2022/04/15 12:00:00+00	1010196	963022.000000	0.000000	14618	0.000000	17428903.000000	0.000000	32556.000000	0.000000
12176	13251	ONTARIO	ON	0	2022/04/15 12:00:00+00	1209041	1163003.000000	0.000000	12606	0.000000	23777414.000000	0.000000	33432.000000	0.000000
12177	13252	NORTHWEST TERRITORIES	NT	0	2022/04/15 12:00:00+00	11136	10975.000000	0.000000	21	0.000000	0.000000	0.000000	139.000000	0.000000
12178	13253	REPATRIATED CDN	RC	0	2022/04/15 12:00:00+00	13	13.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000	0.000000
12179	13254	CANADA	CA	0	2022/04/15 12:00:00+00	3614094	2221502.000000	0.000000	38288	0.000000	58066389.000000	0.000000	75197.000000	0.000000

Figure 3.3: Canada's provinces data.



	OBJECTID	Province	Abbreviation	DailyTotals	SummaryDate	TotalCases	TotalRecovered	DailyRecovered	TotalDeaths	DailyDeaths	TotalTested	DailyTested	TotalActive	DailyActive	TotalHospital
7373	8105	ALBERTA	AB	391	2021-05-30 12:00:00+00:00	227246	216954.000000	787.000000	2219	5.000000	4522783.000000	6609.000000	8073.000000	-401.000000	446.00
7838	8615	ALBERTA	AB	76	2021-06-30 12:00:00+00:00	231987	228631.000000	151.000000	2301	2.000000	4692231.000000	6332.000000	1055.000000	-77.000000	165.00
8288	9065	ALBERTA	AB	187	2021-07-30 12:00:00+00:00	234295	230312.000000	49.000000	2328	3.000000	4866898.000000	8293.000000	1655.000000	135.000000	90.00
8753	9530	ALBERTA	AB	3056	2021-08-30 12:00:00+00:00	252010	238213.000000	1278.000000	2371	7.000000	5107111.000000	29340.000000	11426.000000	1771.000000	401.00
9218	9995	ALBERTA	AB	1706	2021-09-30 12:00:00+00:00	298172	275200.000000	1737.000000	2717	20.000000	5532675.000000	17659.000000	20255.000000	-51.000000	1083.00
9668	10445	ALBERTA	AB	0	2021-10-30 12:00:00+00:00	322989	311738.000000	0.000000	3093	0.000000	5857350.000000	0.000000	8158.000000	0.000000	765.00
10133	10954	ALBERTA	AB	238	2021-11-30 12:00:00+00:00	335247	327454.000000	537.000000	3248	6.000000	6125651.000000	5359.000000	4545.000000	-305.000000	434.00
10583	11478	ALBERTA	AB	0	2021-12-30 12:00:00+00:00	357623	336917.000000	0.000000	3310	0.000000	6374569.000000	0.000000	17396.000000	0.000000	349.00
11048	11943	ALBERTA	AB	0	2022-01-30 12:00:00+00:00	487436	442605.000000	0.000000	3531	0.000000	6728210.000000	0.000000	41300.000000	0.000000	1496.00
11933	13008	ALBERTA	AB	4567	2022-03-30 12:00:00+00:00	540733	0.000000	0.000000	4074	30.000000	6932618.000000	19036.000000	0.000000	0.000000	964.00

Figure 3.4: Collected data for Alberta.

During epidemics and pandemics, infected case rates (ICR) and recovery rates are critical indicators. The continuing coronavirus disease 2019 (COVID-19) pandemic has been visualized in this chapter.

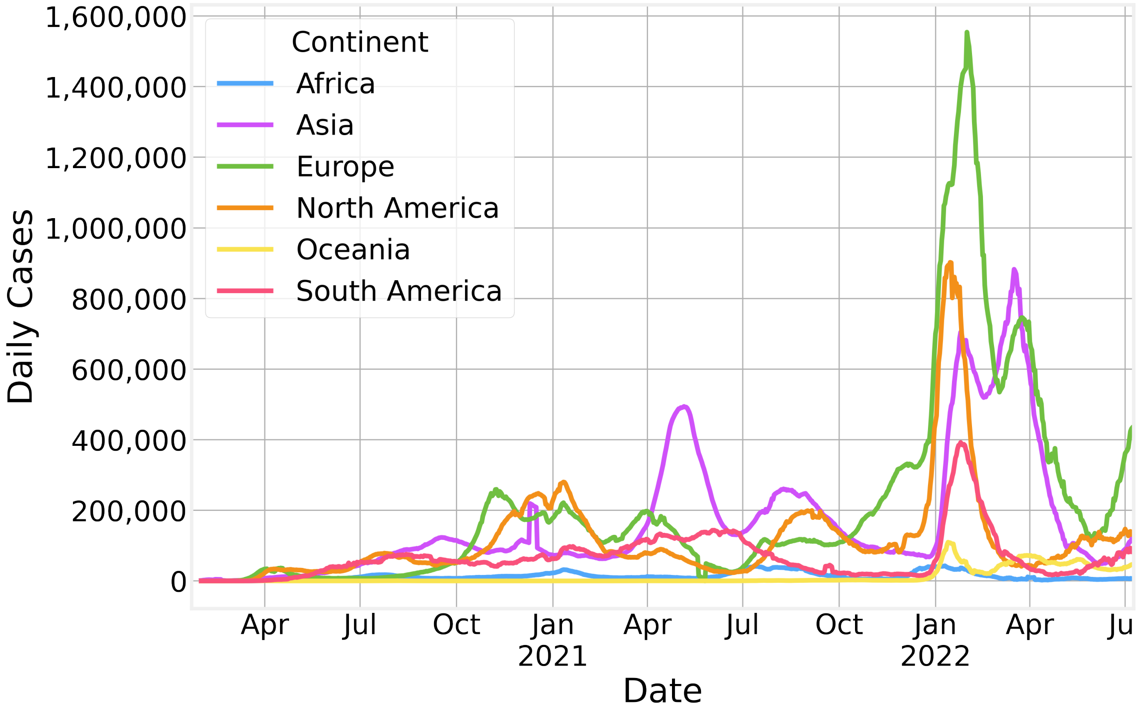


Figure 3.5: Continent Covid-19 daily cases

3.2 Covid-19 in Canada and the world

3.2.1 Countries infected cases

From figure 3.5 we can see that the continent European is leading in the number of confirmed cases, but based on the selected countries, figure 3.7, the United States has the largest COVID-19 epidemic among these 15 countries, whereas Germany, the United Kingdom, and Italy have major COVID-19 epidemics in Europe. On the other hand, and as shown in figure 3.6, it is evident that Denmark and Sweden are the most affected nations when we look at the number of instances broken down by individual.

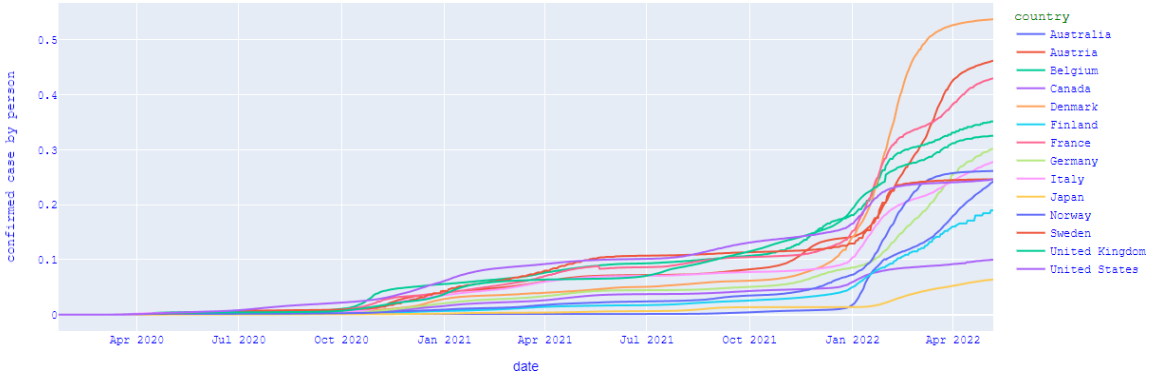


Figure 3.6: Counties Covid-19 total confirmed cases by person

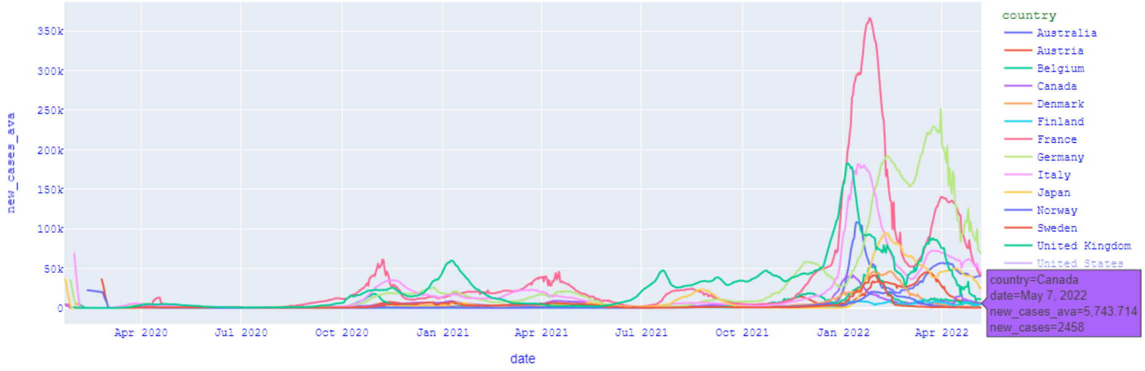
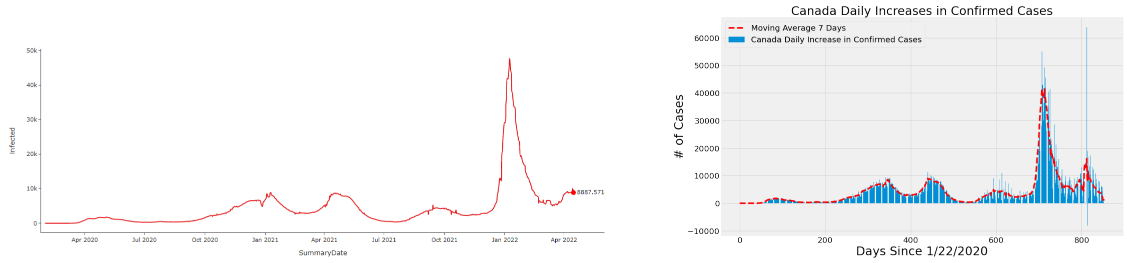


Figure 3.7: Countries Covid-19 daily confirmed cases

3.2.2 How some Countries have brought Covid-19 cases down To early zero.

The number of cases in Canada is nearly the lowest among other countries, which is partly due to the rapid response to the epidemic, as well as the subsequent restrictions and healthcare safeguards. Also If we look at the number of instances through time (figure 3.8), we can see that the country has gone through four key transformations that have prompted decision-makers to make different healthcare decisions over time. However, and if we look to Figure 3.9 we see that South Korea have the lowest number of cases, so what caused this, then?. In 2015, South Korea suffered an outbreak of Middle East respiratory syndrome coronavirus infection. From that time, South Korean hospitals are ready for the next outbreak of contagious illnesses. Respectful preparations were made to healthcare workers, facilities, and the overall system. However, a lot of professionals today believe that the preparations were enough to make the Covid-19 cases almost zero [18].



(a)

(b)

Figure 3.8: Canada Covid-19 Daily cases

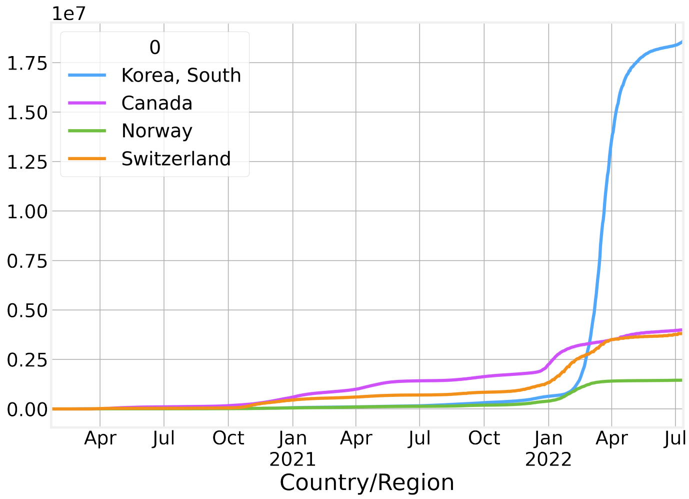


Figure 3.9: Countries with lowest Covid-19 case Vs Canada.

3.2.3 Countries vaccination

Once Fizzier and others released their vaccine, our chosen countries promptly began vaccinating their populations. According to figure 3.10, all countries began administering the Covid-19 vaccine in January 2021, and they are all progressing at the same rate. The fact that all governments prioritize vaccination for their people indicates valid concerns among legislators whose primary responsibility is to their people. The unbalanced distribution of limited vaccine volumes between affluent and poor countries, however, is inequitable and inefficient in the event of a pandemic. A core principle of equity supported by health policy in most OECD nations is allocating finite resources for health care according to need - equitable access according to need. This type of allocation is also cost-effective because it maximizes the overall health benefits that may be obtained from given resources.[19]

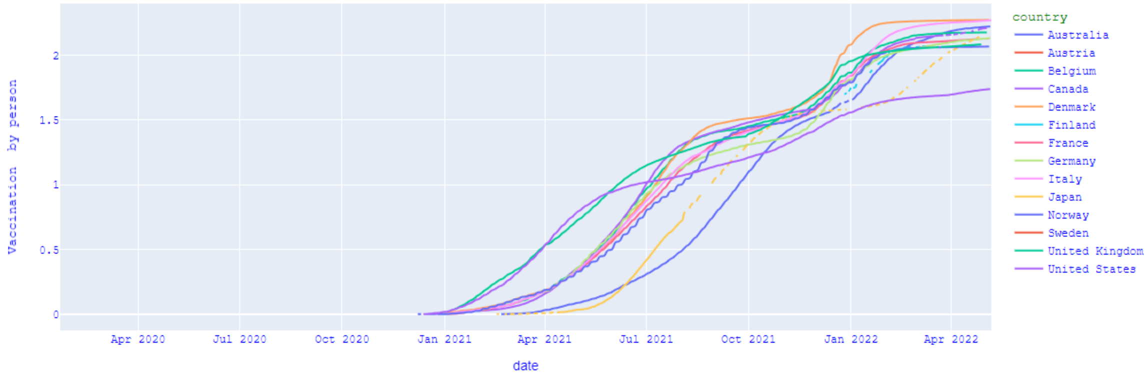
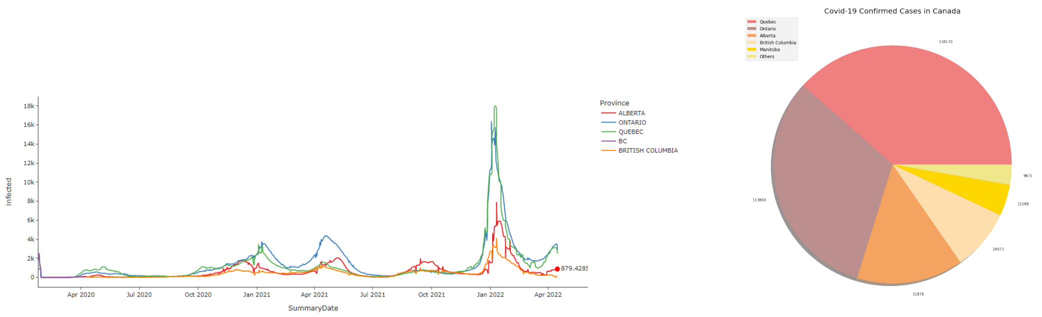


Figure 3.10: Counties vaccination by person

3.3 Covid-19 in Canada's provinces

3.3.1 Provinces infected cases

This section will only cover five provinces, with the remaining provinces being shown in a separate section. Figure 3.11 illustrates that Quebec and Ontario have the most cases in Canada, next to Alberta and British Columbia. The outbreak manifested itself as waves that followed a similar pattern throughout different places but differed in severity. In addition to returned residents and generally physically segregated individuals on cruise ships, exceptions could be reported for very low cumulative cases in particular provinces and regions [20].



(a) (b)

Figure 3.11: Canadian provinces Covid-19 Daily cases

When we view the provinces on the same x-axis,(Figure 3.12) we observe that Alberta has more waves than the other provinces, and its third wave, excluding the very first wave, is stronger than the others. Our careful study of the sash as a difference will be presented in the following parts.

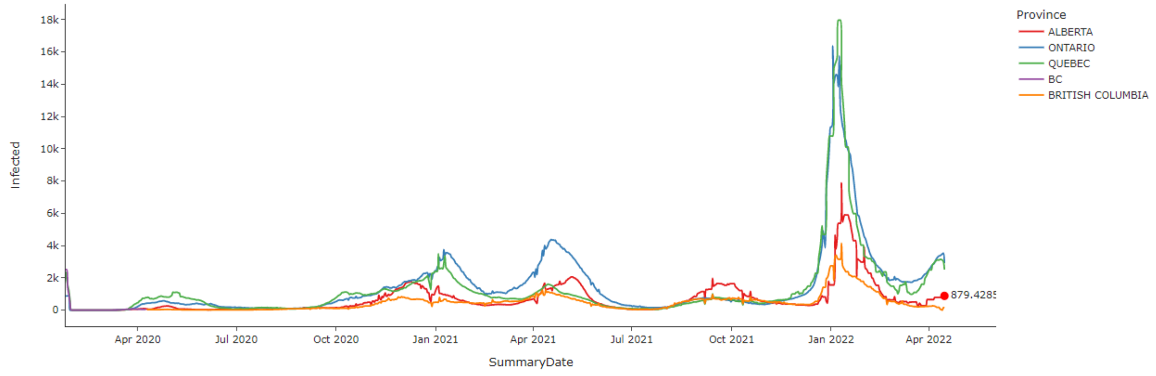
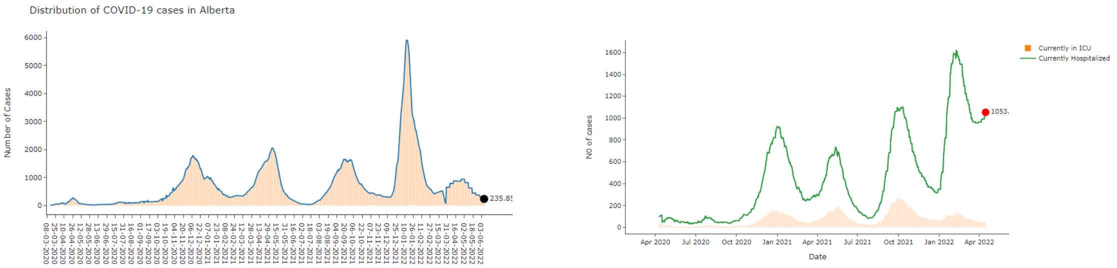


Figure 3.12: Daily cases in Alberta and other provinces in the same line

3.4 Covid-19 in Alberta

3.4.1 Alberta confirmed, hospitalized and ICU cases

A simple visual representation in figure 3.13(a) of Alberta's daily cases is a useful place to start when analyzing instances in this jurisdiction. Even though the number of cases in the tree wave between October 2020 and November 2021 is nearly identical, the hospitalization and ICU cases in figure 3.13(b) suggest the reverse. The third waver, which was essentially non-existent in the other provinces, accounts for the majority of ICU and hospitalization cases. Unlike Ontario, which had three times the population yet fared far better in the fourth wave despite maintaining numerous public health measures in place, Alberta fought vaccine passports, loosened mask laws, and even planned to abandon test, trace, and isolate protocols before backtracking when cases increased [21].



(a) Alberta confirmed cases

(b) ICU and Hospitalized cases

Figure 3.13: Covid-19 in Alberta

According to [21], Dr. Ilan Schwartz, a physician and assistant professor of infectious diseases at the University of Alberta in Edmonton, stated that the Alberta government abdicated its obligation to guarantee the health and well-being of individuals in the fourth wave. Alberta took a risk by removing all restrictions and declaring the pandemic to be over. Jason Kenney famously declared that we were no longer in the post-pandemic age, that COVID was no longer a threat, and threw caution to the wind. However, what made things a lot worse was the inability to respond to statistics that showed an increase in the number of cases and a meticulous examination of the reproduction number, which was always more than one.

All of the preceding points are the main emphasis of the upcoming chapters, which will delve deeper into analyzing when and how decision-makers should react and make the best decision possible.

Chapter 4

Epidemiological Modelling for COVID-19 Disease Prediction in Canada.

4.1 Introduction

Policy decisions in health care often have to be made despite an incomplete understanding of how interactions between agent, environment, and host-level factors affect infection transmission and illness progression. Epidemiological disease models combine existing information from the field and experimental investigations with expert opinion to obtain insight into the dynamics of infection and disease control, allowing them to handle these issues [22].

4.1.1 What are epidemiological models?

A model is a depiction of a physical process or system created to help people appreciate and understand it better. Models are created to better comprehend the impact of external influences on outputs by representing the interactions between the system's components and expressing ideas about the system's behaviour [22]. Epidemiological models are typically defined as mathematical and/or logical representations of disease transmission epidemiology and associated processes. In the context of animal disease management,'models' can be defined more generally to include a variety of statistical/mathematical methods that consider factors other than disease propagation[22].

4.1.2 Epidemiological models with covid-19 data

The study by Davies et al reported in [23] is based on compartmental modelling, in which individuals are divided into groups based on their infection or symptom state [3]. The Susceptible-Infectious-Removed (SIR) model is the most famous epidemiologic model.

What is the SIR model?

The SIR model is a compartmental model that describes the dynamics of infectious disease. The compartmental model gets its name from the fact that it divides the population into segments. Each compartment should have the same features. SIR represents the model's three compartments [24].

- Susceptible
- Infectious
- Recovered

Susceptible persons are those who are at risk of becoming infected if they come into contact with infectious people. When the infection occurs, they might be patient. Infectious persons are represented by the infectious group. They can spread the disease to others who are susceptible, and they can recover in a set amount of time. People who have recovered get immunity, which means they are no longer susceptible to the same ailment. The SIR model is a framework for defining how the population of every category can change and evolve [24].

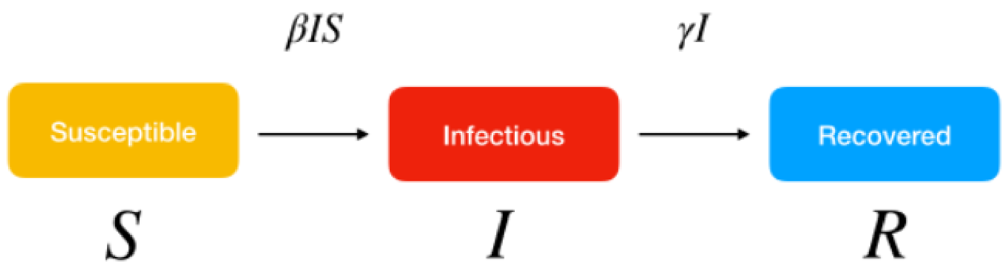


Figure 4.1: SIR equation [24]

Using the SIR model, we can use an ordinary differential equation to describe the number of persons in each compartment.

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

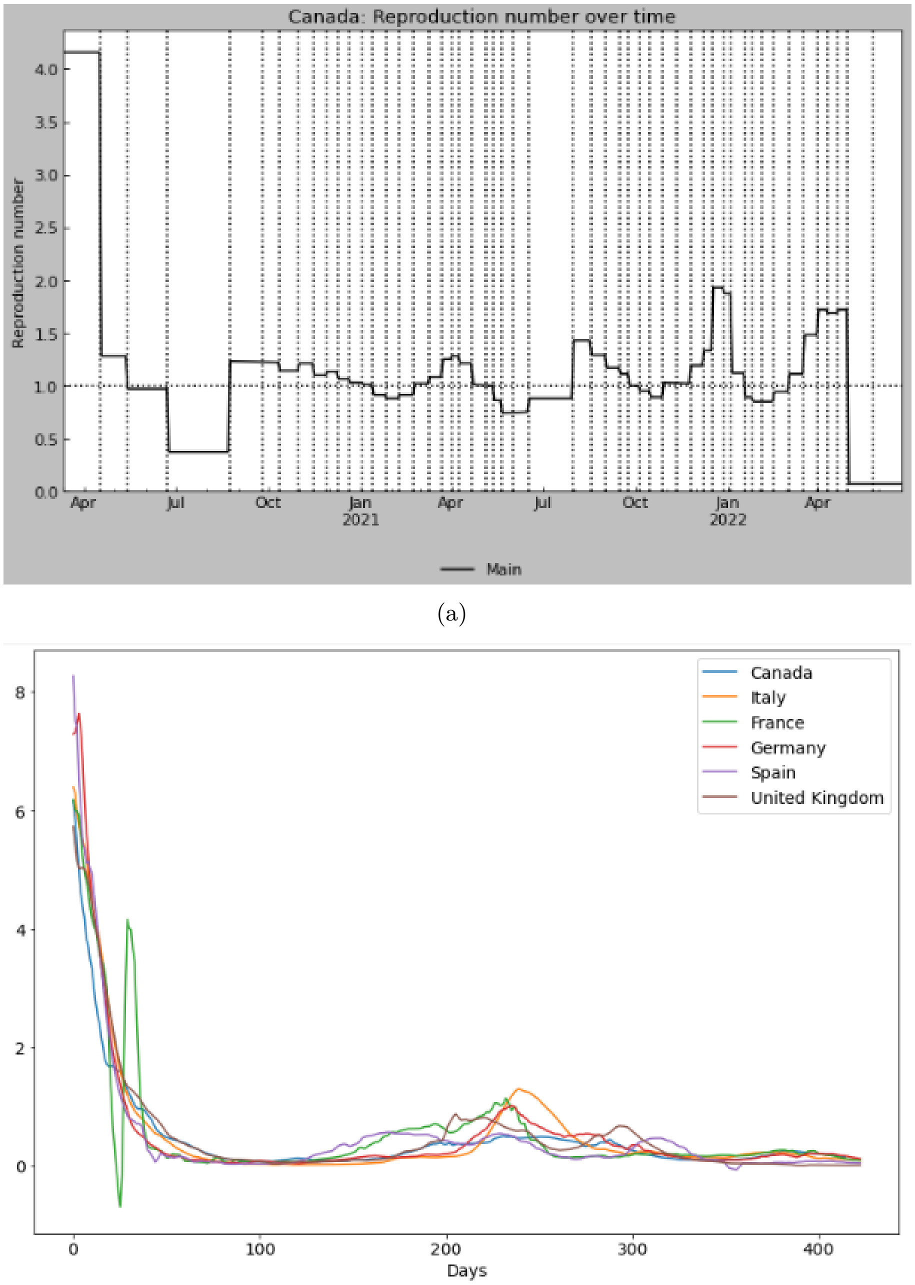
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

Where β is the contagion rate of the pathogen and γ is the recovery rate.
There are various insights gained from being able to estimate the two values:
If D is the average number of days it takes to recover from an infectious disease, it is derived from $D = 1/\gamma$.
Also, we can estimate the nature of the disease in terms of the power of infection $R_0 = \beta/\gamma$.
R_0 is a basic reproduction number that represents the average number of people infected by one another. If it's high, the chances of a pandemic are also high. It's also used to calculate the herd immunity threshold (HIT). The balanced state is shown by the fundamental reproduction number multiplied by the fraction of non-immune people (susceptible). The number of infected persons is always increasing [24].

4.2 Canada S-I-R trend analysis

Some countries attracted the early notice of the epidemic due to their drastically lower mortality rate than other European countries at the time, as compared to Canada. Germany's higher testing rate, ability to ramp up testing more swiftly and earlier than many of its EU competitors, having more ICU beds, and younger persons becoming infected were all factors. All of that changed with the second wave, which hit Europe in the fall and lasted well into the winter. In mid-December 2020, the German government enforced a tight lockdown, which was extended numerous times. More than 1,700 deaths were documented in a single day at its worst point in January 2021, with long-term care homes being the hardest hit [25].

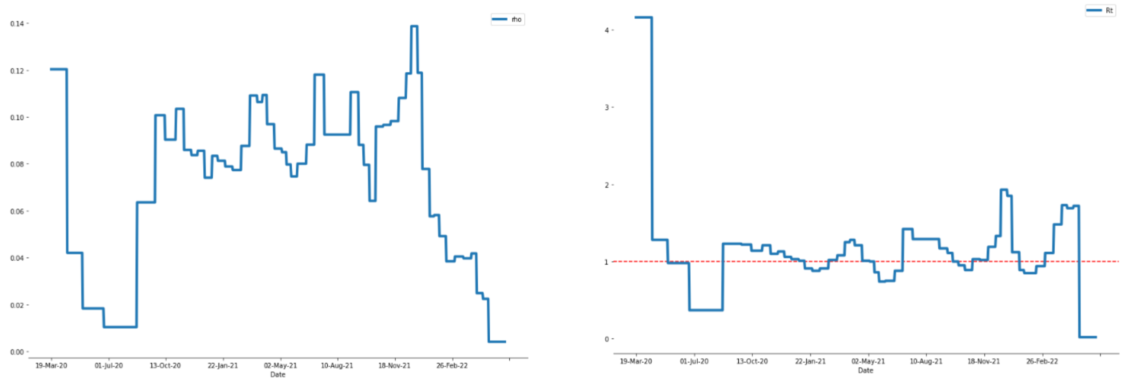


(a)

(b)

Figure 4.2: SIR for Canada and other countries

The competition and graph of the SIR trend, figure 4.2, and figure 4.3, demonstrate that there were five times where the reproduction number was larger than one, and now we can understand why Canada entered 2022 with a record number of cases attributable to Covid-19. The country's number of cases hit new highs in December, and Canada's senior public health officer, Dr. Theresa Tam, said Omicron has "rapidly" displaced Delta as the dominant strain in the two-year-long pandemic. It prompted new restrictions in some regions, disrupted the resumption of school after the Christmas break, and pushed enterprises to dramatically decrease capacity or close entirely [25].
The spread of Covid-19 and its variants, which peaked in January 2021 and again in mid-April 2021, when it reported more than a million cases a little over a year after the WHO formally declared a pandemic, was Canada's fourth wave's worst because of the poor interpretation of the epidemiological models. By mid-May 2021, Canada had crossed the 25,000-death mark [25].



(a)

(b)

Figure 4.3: SIR for Canada

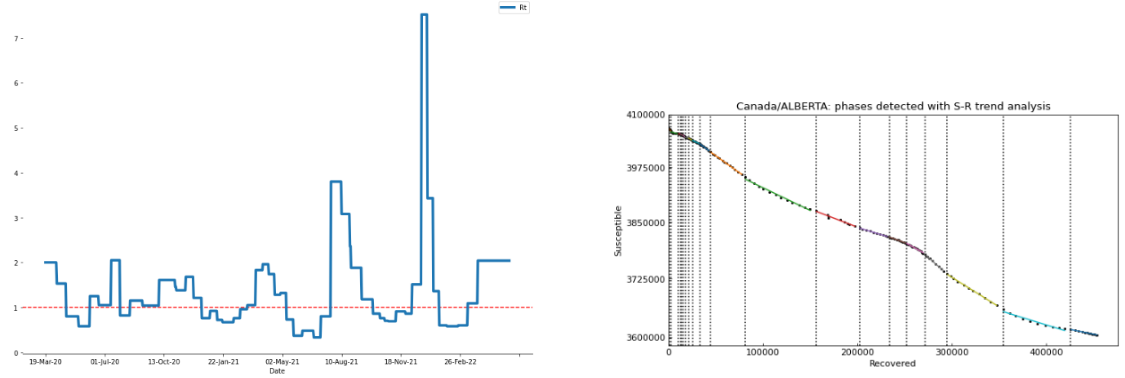
Long-term care homes, particularly in Alberta, Ontario and Quebec, were ill-prepared and disproportionately affected during the pandemic's first wave. During the first wave, breakouts at hundreds of these establishments were responsible for more than 80% of all deaths [25].

During the period time, the average effective reproduction numbers in nine Canadian provinces were more than one, and non-pharmaceutical interventions (NPIs) in Ontario and Saskatchewan had minimal impact on the dynamics of COVID-19 epidemics. The average infection probability in Alberta reached its greatest level more than once since the start of the COVID-19 pandemic in Canada [26].

4.2.1 Alberta S-I-R trend analysis

In this section, we describe the findings of a study that used statistical models to investigate the characteristics of COVID-19 in Alberta. The main goal of this section is to show how the SIR modelling methodology can be used to interpret COVID-19 data. We anticipate that the research will help us better understand COVID-19's complicated characteristics and development in Alberta. Readers are encouraged to pay attention to the reproduction number trend while interpreting the data.

To investigate the spread of the COVID-19 outbreak in Alberta, we plotted reproduction number rates in the province from Mars 19, 2020 to July 13, 2022. The figure 4.4 demonstrates that from May 2020, the average effective reproduction numbers in Alberta were more than nine-time and less than five-time, indicating that the COVID-19 epidemic has not been controlled effectively. Even if Alberta's government responded quickly to the COVID-19 epidemic, putting in place robust public health containment measures within three weeks of finding the first positive case in the province (Mar. 5, 2020), and during this time, all patients presenting to the hospital were also checked for core respiratory symptoms [27], The government quickly lost control of the outbreak.



(a)

(b)

Figure 4.4: Alberta reproductive number

Chapter 5

Machine Learning-based COVID-19 Prediction for Canada.

5.1 Prediction with Machine Learning

In the context of machine learning, a prediction is an information output that results from the input of data and the execution of an algorithm. The primary challenge with any prediction method is that training data, the inputs you'll need to start generating good results, must either be developed (by employing experts to classify things, for example) or obtained from existing sources (say, health records). Some data can be easily obtained from public sources (think of weather and map information). Consumers may also willingly provide personal data if they believe it will benefit them [28].

There is no one-size-fits-all machine learning algorithm for every problem, and this is especially true for supervised learning (i.e. predictive modelling). However, all supervised machine learning methods for predictive modelling are based on the same idea. Learning a target function (f) that best maps input variables (X) to an output variable (Y) is how machine learning methods are described: $f = Y(X)$

This is a general learning task in which we want to make future predictions (Y) based on new examples of input variables (X). We have no idea what the function (f) looks like or in what shape it takes. If we did, we wouldn't need to learn it from data using machine learning methods because we'd be able to apply it right away. Learning the mapping Y = f(X) to produce Y predictions for new X is the most prevalent sort of machine learning. Our goal is to make the most accurate forecasts possible, which is known as predictive modelling or predictive analytics. [29].

5.1.1 Alberta Data

In this section, we are working with confirmed cases in Alberta. First let's look at the distribution of the data Figure 5.1, and start our predictions.

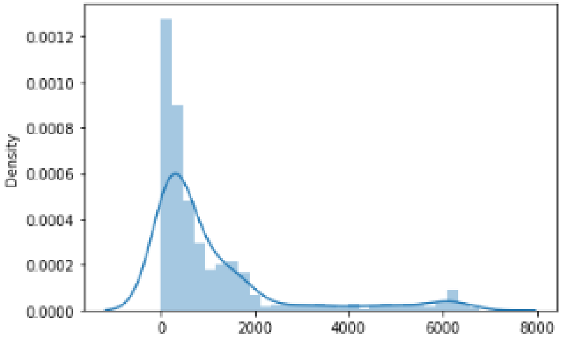
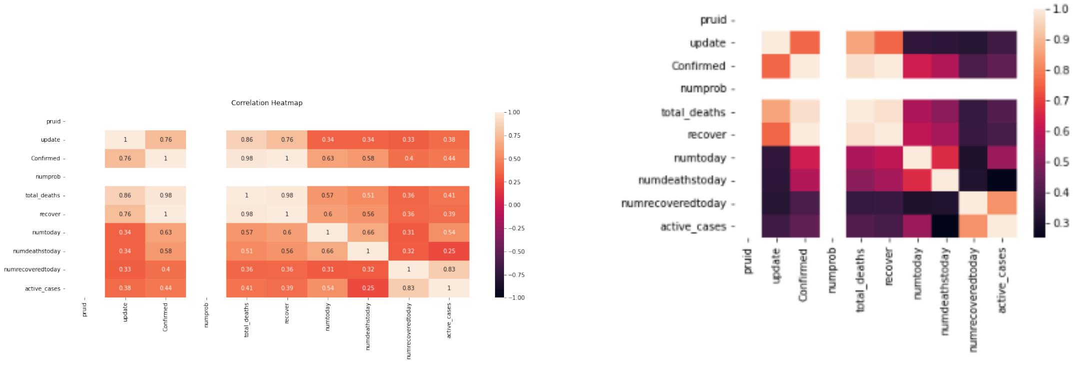


Figure 5.1: Confirmed cases distribution

5.2 Seaborn heatmap

To begin, we'll look at a Seaborn data visualization. It provides a way to present data in a statistical graph format that is both instructive and appealing to the eye. A heatmap is one of the seaborn components that depicts variations in linked data using a colour palette. Figure 5.2 primarily focuses on a correlation heatmap and how it is generated for an Alberta dataframe using seaborn in conjunction with pandas and matplotlib.

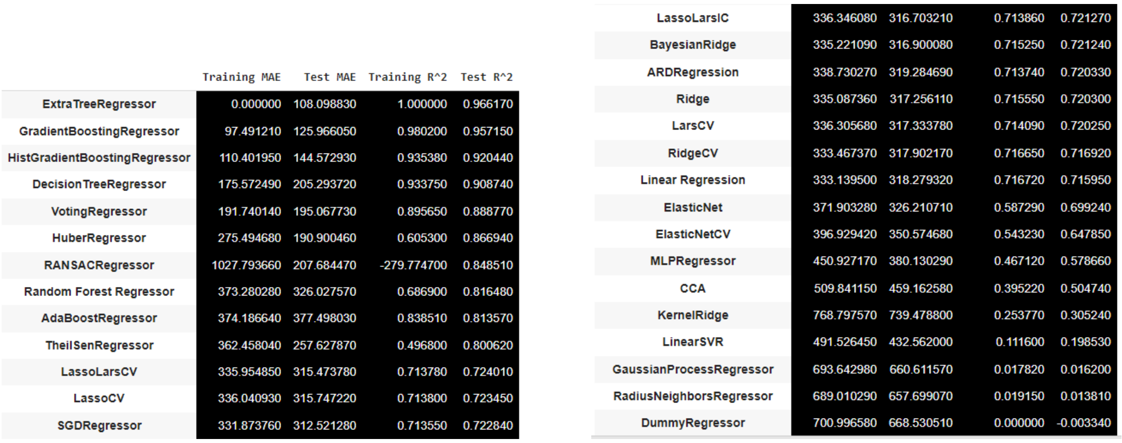


(a) (b)

Figure 5.2: Alberta Covid-19 seaborn correlation heatmap

5.3 Finding the Most Appropriate Distribution for Our Data

The prediction model for Covid-19 daily cases is built using machine learning algorithms such as Randomforest, Ridge Regression, Lasso Regression, and ElasticNet Regression. We also compared these methods using the RMSE parameter. Finally, we combined the results in a table to improve the accuracy of our model.
This process utilizes the X and Y matrices from the Train and Test sets as input, and it applies them to all of the Classifiers in the dict classifier. Typically, training the SVM, Random Forest, and Gradient Boosting Regressor take a long time. As a result, it's better to start by training them on a smaller dataset and then comment them out depending on the test accuracy score.



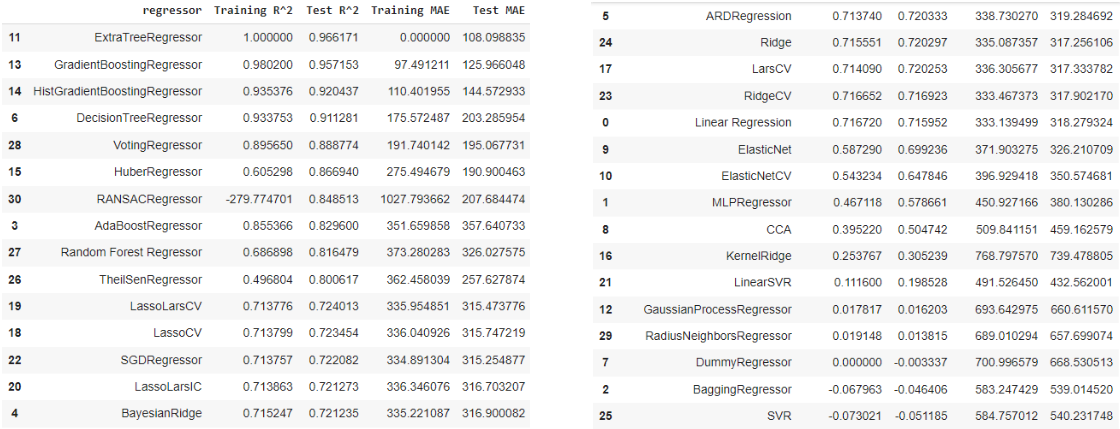
	Training MAE	Test MAE	Training R^2	Test R^2
ExtraTreeRegressor	0.000000	108.098830	1.000000	0.966170
GradientBoostingRegressor	97.491210	125.966050	0.980200	0.957150
HistGradientBoostingRegressor	110.401950	144.572930	0.935380	0.920440
DecisionTreeRegressor	175.572490	205.293720	0.933750	0.908740
VotingRegressor	191.740140	195.067730	0.895650	0.888770
HuberRegressor	275.494680	190.900460	0.605300	0.866940
RANSACRegressor	1027.793660	207.684470	-279.774700	0.848510
Random Forest Regressor	373.280280	326.027570	0.686900	0.816480
AdaBoostRegressor	374.186640	377.498030	0.838510	0.813570
TheilSenRegressor	362.458040	257.627870	0.496800	0.800620
LassoLarsCV	335.954850	315.473780	0.713780	0.724010
LassoCV	336.040930	315.747220	0.713800	0.723450
SGDRegressor	331.873760	312.521280	0.713550	0.722840

(a)

LassoLarsIC	336.346080	316.703210	0.713860	0.721270
BayesianRidge	335.221090	316.900080	0.715250	0.721240
ARDRegression	338.730270	319.284690	0.713740	0.720330
Ridge	335.087360	317.256110	0.715550	0.720300
LarsCV	336.305680	317.333780	0.714090	0.720250
RidgeCV	333.467370	317.902170	0.716650	0.716920
Linear Regression	333.139500	318.279320	0.716720	0.715950
ElasticNet	371.903280	326.210710	0.587290	0.699240
ElasticNetCV	396.929420	350.574680	0.543230	0.647850
MLPRegressor	450.927170	380.130290	0.467120	0.578660
CCA	509.841150	459.162580	0.395220	0.504740
KernelRidge	768.797570	739.478800	0.253770	0.305240
LinearSVR	491.526450	432.562000	0.111600	0.198530
GaussianProcessRegressor	693.642980	660.611570	0.017820	0.016200
RadiusNeighborsRegressor	689.010290	657.699070	0.019150	0.013810
DummyRegressor	700.996580	668.530510	0.000000	-0.003340

(b)

Figure 5.3: scores of each model$_files$



	regressor	Training R^2	Test R^2	Training MAE	Test MAE
11	ExtraTreeRegressor	1.000000	0.966171	0.000000	108.098835
13	GradientBoostingRegressor	0.980200	0.957153	97.491211	125.966048
14	HistGradientBoostingRegressor	0.935376	0.920437	110.401955	144.572933
6	DecisionTreeRegressor	0.933753	0.911281	175.572487	203.285954
28	VotingRegressor	0.895650	0.888774	191.740142	195.067731
15	HuberRegressor	0.605298	0.866940	275.494679	190.900463
30	RANSACRegressor	-279.774701	0.848513	1027.793662	207.684474
3	AdaBoostRegressor	0.855366	0.829600	351.659858	357.640733
27	Random Forest Regressor	0.686898	0.816479	373.280283	326.027575
26	TheilSenRegressor	0.496804	0.800617	362.458039	257.627874
19	LassoLarsCV	0.713776	0.724013	335.954851	315.473776
18	LassoCV	0.713799	0.723454	336.040926	315.747219
22	SGDRegressor	0.713757	0.722082	334.891304	315.254877
20	LassoLarsIC	0.713863	0.721273	336.346076	316.703207
4	BayesianRidge	0.715247	0.721235	335.221087	316.900082

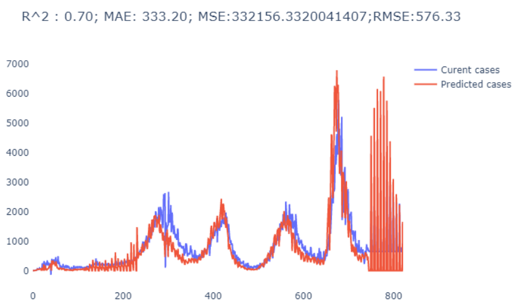
(a)

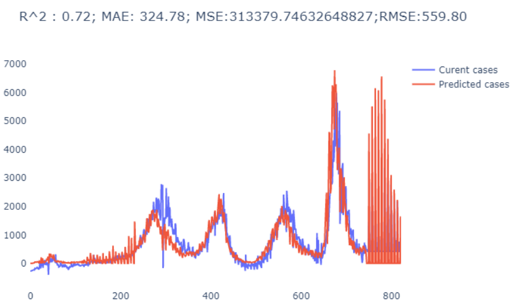
5	ARDRegression	0.713740	0.720333	338.730270	319.284692
24	Ridge	0.715551	0.720297	335.087357	317.256106
17	LarsCV	0.714090	0.720253	336.305677	317.333782
23	RidgeCV	0.716652	0.716923	333.467373	317.902170
0	Linear Regression	0.716720	0.715952	333.139499	318.279324
9	ElasticNet	0.587290	0.699236	371.903275	326.210709
10	ElasticNetCV	0.543234	0.647846	396.929418	350.574681
1	MLPRegressor	0.467118	0.578661	450.927166	380.130286
8	CCA	0.395220	0.504742	509.841151	459.162579
16	KernelRidge	0.253767	0.305239	768.797570	739.478805
21	LinearSVR	0.111600	0.198528	491.526450	432.562001
12	GaussianProcessRegressor	0.017817	0.016203	693.642975	660.611570
29	RadiusNeighborsRegressor	0.019148	0.013815	689.010294	657.699074
7	DummyRegressor	0.000000	-0.003337	700.996579	668.530513
2	BaggingRegressor	-0.067963	-0.046406	583.247429	539.014520
25	SVR	-0.073021	-0.051185	584.757012	540.231748

(b)

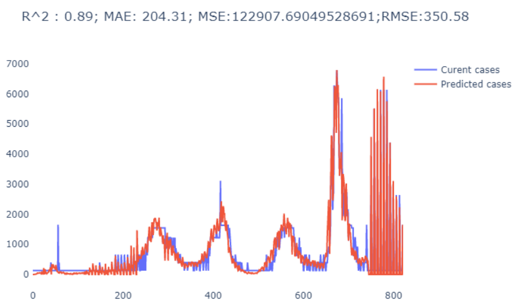
Figure 5.4: scores of each model$_files$

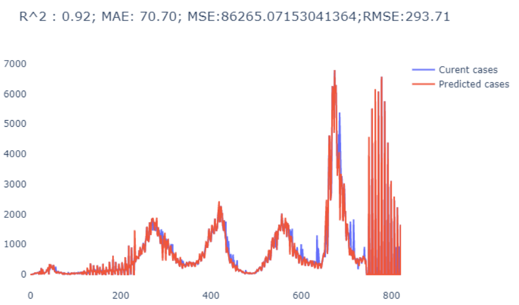
- Kernel Ridge prediction Bayesian Ridge prediction



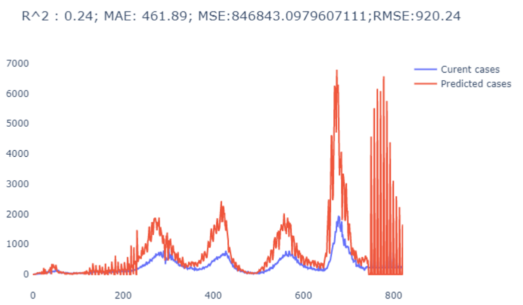


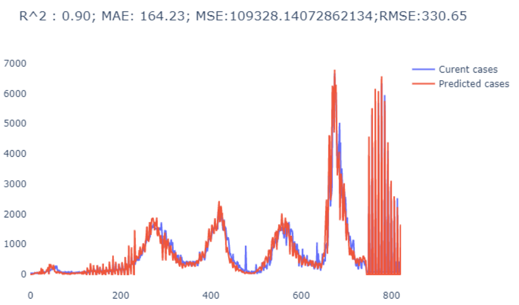
- Decision Tree Regressor prediction Extra Trees Regressor prediction





- Linear SVR prediction Gradient Boosting Regressor prediction





5.4 Classification and Regression Trees

Decision Trees are a common sort of machine learning algorithm for predictive modelling, and its algorithm is part of the supervised learning algorithms family.

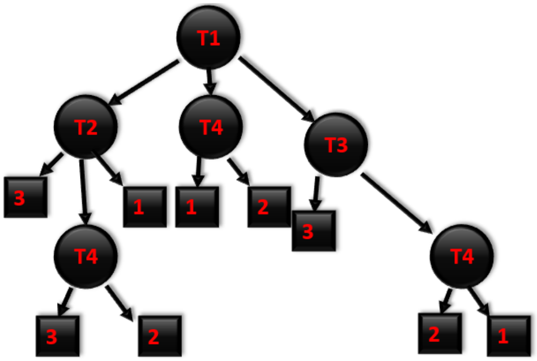
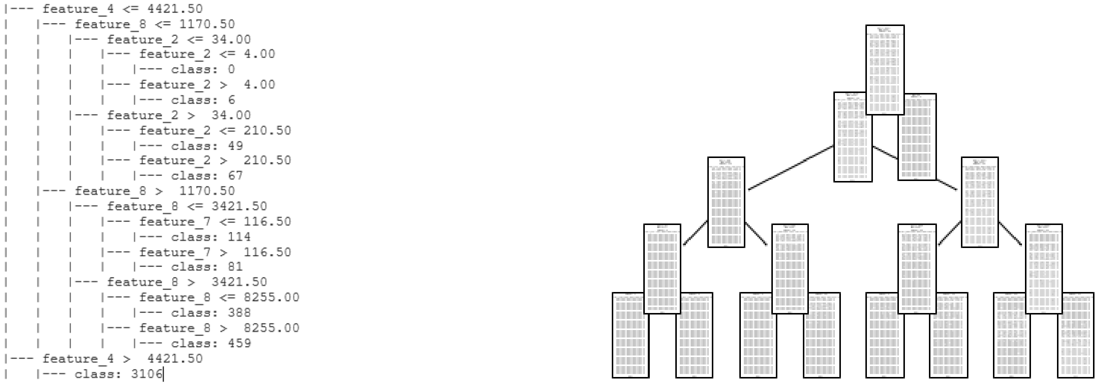


Figure 5.5: decision tree example

A decision tree is a tree with tests as the inner nodes and categories as the leaf nodes. Figure 5.5 depicts one example. By filtering an input pattern through the tree's tests, a decision tree assigns a class number to it. Each test yields data that are both thorough and mutually exclusive. T2 in Figure 5.5 has three results: the one on the far left assigns the input pattern to class 3, the one in the centre sends it down to test T4, and the one on the far right assigns it to class 1. Leaf nodes are represented by their class number, as is customary [30].



```
|--- feature_4 <= 4421.50
|   |--- feature_8 <= 1170.50
|   |   |--- feature_2 <= 34.00
|   |   |   |--- feature_2 <= 4.00
|   |   |   |   |--- class: 0
|   |   |   |--- feature_2 >  4.00
|   |   |   |   |--- class: 6
|   |   |--- feature_2 >  34.00
|   |   |   |--- feature_2 <= 210.50
|   |   |   |   |--- class: 49
|   |   |   |--- feature_2 >  210.50
|   |   |   |   |--- class: 67
|   |--- feature_8 >  1170.50
|   |   |--- feature_8 <= 3421.50
|   |   |   |--- feature_7 <= 116.50
|   |   |   |   |--- class: 114
|   |   |   |--- feature_7 >  116.50
|   |   |   |   |--- class: 81
|   |   |--- feature_8 >  3421.50
|   |   |   |--- feature_8 <= 8255.00
|   |   |   |   |--- class: 388
|   |   |   |--- feature_8 >  8255.00
|   |   |   |   |--- class: 459
|--- feature_4 >  4421.50
|   |--- class: 3106
```

(a)

(b)

Figure 5.6: Covid-19 infected cases decision tree

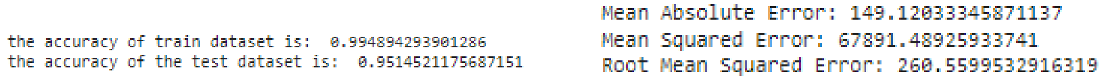
The application of the decision tree approach, unlike most other supervised learning algorithms, may also be employed to address regression and classification questions.

By learning simple decision rules inferred from prior data, we classified Covid-19 daily infected cases for Alberta based on the Sum of Product (SOP) representation. Disjunctive Normal Form is another name for the Sum of Product (SOP). Every branch from the tree's root to a leaf node with the same class is a conjunction (product) of values, while distinct branches terminating in that class constitute a disjunction (sum) [31].

5.5 Random Forest Regression

RF is a regression approach that classifies or predicts the value of a variable by combining the results of many DT algorithms. When RF gets a (x) input vector containing the values of the many evidentiary characteristics investigated for a specific training area, it constructs a number of K regression trees and averages the findings. RF boosts the variety of the trees by making them grow from distinct training data subsets provided by a method called bagging. This prevents the trees from being correlated. Bagging is a training data production strategy that involves resampling the original dataset at random with replacement, i.e., without deleting the data selected from the input sample for the next subset [32].

We used sklearn's RandomForestClassifier module to train our Alberta dataset with 100 estimators, starting at state 0, to produce our confirmed case prediction.



```
the accuracy of train dataset is:  0.994894293901286
the accuracy of the test dataset is:  0.9514521175687151
```

(a)

```
Mean Absolute Error: 149.12033345871137
Mean Squared Error: 67891.48925933741
Root Mean Squared Error: 260.5599532916319
```

(b)

Figure 5.7: Prediction accuracy, RMS, MSE, and MAE

The samples that were not chosen for the training of the k-th tree during the bagging procedure are grouped as part of an out-of-bag subset (oob). The k-th-tree can use these oob elements to measure performance. Without employing an external data subset, RF may obtain an unbiased assessment of the generalization error. As the number of trees grows, the generalization error decreases, indicating that the RF does not overfit the input [32].

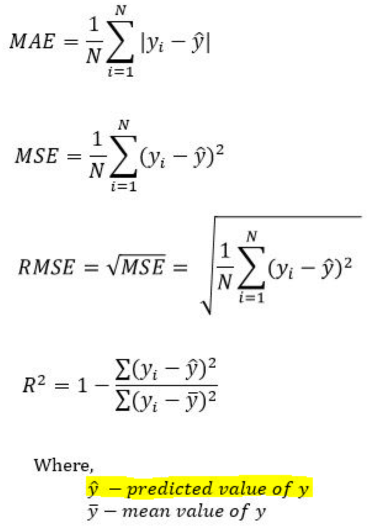
5.6 Alberta Covid-19 prediction using Gaussian naïve Bayes

Here, we'll talk about the Gaussian Naive Bayes classifier. Naive Bayes is one of the most well-known machine learning algorithms. This is an example of a potential classification algorithm. We created a Python code to measure the accuracy of this model. The conditional probability formula is as follows:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{12}$$

, and we acquire Gaussian Naive Bayes model accuracy with this programme (%).

To see how well is the prediction, we will calculate the difference between $y - y_p redicted$ or $(y - y_p redicted)^2$, and the following equations will be generated:



$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ $-$ *predicted value of y*

$\bar{y}$ $-$ *mean value of y*

- MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaging the absolute difference over the data set.
- MSE (Mean Squared Error) represents the difference between the original and predicted values extracted by squaring the average difference over the data set.
- $RMSE$ (Root Mean Squared Error) is the error rate by the square root of MSE.
- R^2 (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 are interpreted as percentages. The higher the value is, the better the model is.

The algorithm applied in this section is as follows;

Algorithm 1 Pseudocode of naïve bayes algorithm

```
Input: Training dataset T,
$F = (f_1, f_2, f_3, ...., f_n)$ is the value of predicted variable in testing dataset.
Output: A class of testingset.
Steps:
  • Read the training sataset T
  • Choose your test size to split between training and testing sets
  • Create a model applying the Gaussian naivebayes
  • Fit $X_train, y_train$ to the model
  • calculate the model score of $X_train, y_train$
  • predict the daily number of cases 'numtoday'
Graph the output and calculate the MAE, MSE, RMSE, and R-Squared.
```

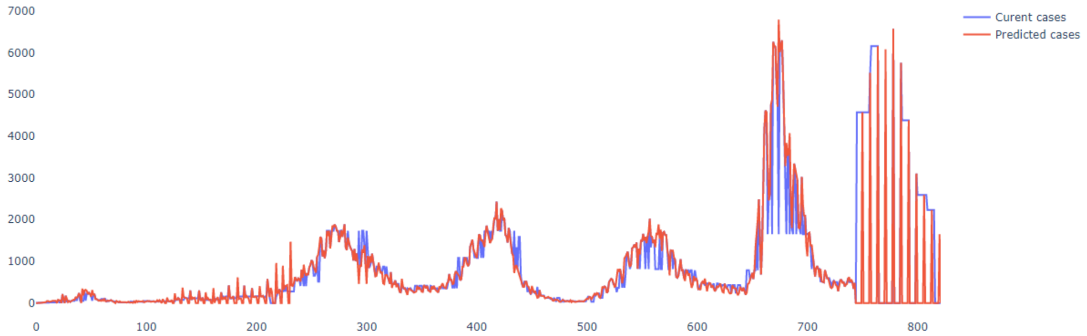


Figure 5.8: Current vs predicted daily cases

From the table below, we can see that the error rate between the current cases and the predicted cases is 0.14.

Metric	Alberta daily cases
MAE	267.37
MSE	968025.42
RMSE	983.88
R-Squared	0.14

Chapter 6

Bayesian Analysis for COVID-19 Prediction.

6.1 Introduction to Bayesian Analysis

6.1.1 Bayesian inference

Bayesian inference is the process of fitting a probability model to a set of data and summarising the outcome using a probability distribution on model parameters and unobserved quantities such as predictions for additional observations [33].

6.1.2 Probability and inference

The explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis is a key feature of Bayesian approaches. The following three steps can be used to summarise the Bayesian data analysis process [33]:

- Create a full probability model, which is a probability distribution that includes all observable and unobservable variables in an issue. The model should be in line with what we know about the underlying scientific topic and how data is collected.

- Using observed data to condition: computing and interpreting the relevant posterior distribution—the conditional probability distribution of the unobserved quantities of ultimate interest given the observed data.

- Examining the model's fit and the implications of the posterior distribution that arises: how well does the model fit the data, are the substantive findings plausible, and how sensitive are the results to the modelling assumptions in step 1? As a result, the model can be changed or expanded, and the three processes can be repeated.

Bayes' theorem

The ability to comprehend statistical inferences reasonably is a fundamental motivation for Bayesian analysis which have the goal of extracting inferences from numerical data. Testing Covid-19 in a broad group of patients, for example, is neither possible nor ethical. As a result, judgments about genuine probabilities (of infection) and, in particular, disparities between them, must be established on a sample of patients. Probabilities are numbers in the range [0, 1], with both extremes included. The product rule is one of the rules that govern probabilities [34]:

$$p(H, D) = p(H/D).p(D)$$

and this is how we read it: The likelihood of D and H is the chance of H given D multiplied by the probability of D. This can also be written in the following format:

$$p(D, H) = p(D/H).p(H)$$

We can write the following if the terms on the left are equal:

$$p(D/H).p(H) = p(H/D).p(D)$$

And if we reorder it, we get **Bayes' theorem** [34]:

$$p(H/D) = \frac{p(D/H).p(H)}{p(D)} \tag{6.1}$$

with:

- $p(H)$: Prior distribution should reflect what we know about the value of some parameter before seeing the data D
- $p(D|H)$: Likelihood is how we will introduce data in our analysis. It is an expression of the plausibility of the data given the parameters.
- $p(H|D)$: Posterior distribution is the result of the Bayesian analysis and reflects all that we know about a problem (given our data and model).
- $p(D)$: Evidence also known as marginal likelihood. Formally, the evidence is the probability of observing the data averaged over all the possible values the parameters can take.

6.2 Bayesian Analysis for COVID-19 Prediction

6.2.1 Markov Chain Monte Carlo (MCMC)

PyMC3, and more generally Markov Chain Monte Carlo (MCMC), is the first library used in this section to analyze the data for Canada. MCMC is performed using a huge number of algorithms. The majority of these algorithms can be summarised as follows:

Algorithm 2 MCMC algorithm

```
Require: Begin at the present location.
Require: Suggest a new position
Require: Accept or reject the new position depending on how well it follows the data and previous
  distributions.

  if you accept then
      proceed to your new position.
      Return to the first step.

  else if  then Return to the first step.
  end if
Require: After a large number of iterations, return all accepted positions.
```

Notice that only the current location matters in the pseudocode for the algorithm above (new positions are investigated only near the current position). This trait is known as memorylessness, which means that the algorithm doesn't care how it got to its current position; all it cares about is that it's there [35].

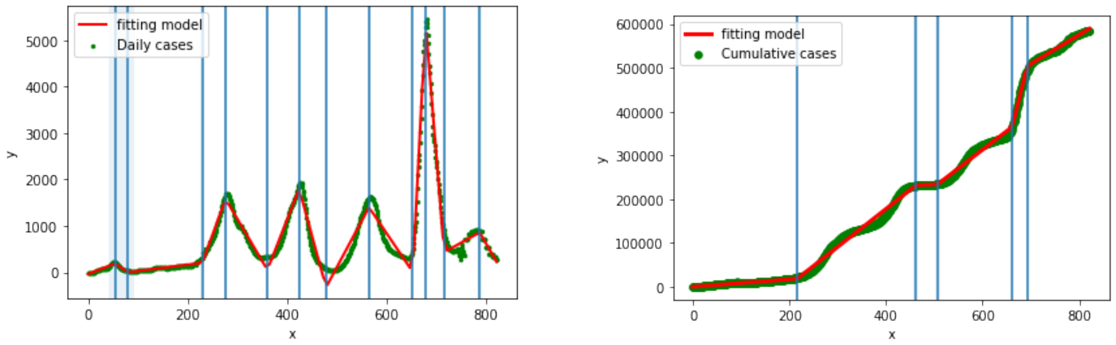
6.3 Piecewise-regression (aka segmented regression)

Piecewise regression, also known as broken-line regression, is a type of segmented regression in which a linear regression model is fitted to data with one or more breakpoints where the gradient changes. The piecewise-regression Python module employs Muggeo's [36] technique, in which the breakpoint positions and straight-line models are both fitted using an iterative process. This user-friendly tool contains an automatic statistical analysis that provides confidence intervals for all model variables as well as hypothesis testing for the presence of breakpoints.

Fitting a continuous straight line model to data that includes some changes in gradient, known as breakpoints, is a typical challenge in many domains, and investigating Covid-19 daily cases is one example.

The global difficulty of estimating breakpoint positions and the local problem of fitting line segments given breakpoints are both involved in fitting such models. Using linear regression to fit line segments together and a global optimization technique to discover

breakpoints are two possible ways. Alternatively, using scipy, we may use a nonlinear least-squares technique [37].
Muggeo [38] developed an alternate method that involves fitting the breakpoint positions and line segment models simultaneously using an iterative process, which is computationally efficient and allows for robust statistical analysis. This approach is implemented in a number of R packages, including Muggeo's own segmented R package [39]. However, there were no similar resources in Python prior to the piecewise-regression module. Figure 6.1 depicts an example plot. A model was fit to the data after it was generated with 13 breakpoints and some noise. The maximum likelihood estimators for straight line segments and breakpoint places are shown in the graph.



(a)

(b)

Figure 6.1: Covid-19 infected cases Piecewise Regression

6.3.1 Piecewise Regression Mathematical application

Muggeo's derivation is followed here [38].
The principle behind piecewise linear regression is that if the data follows various linear trends in different parts of the data, the regression function should be modelled in "pieces." The equations that make up our problem are listed below:

$$\text{f(x)} = \begin{cases} \alpha_1 x + c + \beta_1(x - \Psi_1)H(x - \Psi_1) + \zeta & \text{if } \psi_1 < x \leqslant \psi_2 \\ \alpha_2 x + c + \beta_2(x - \Psi_2)H(x - \Psi_2) + \zeta & \text{if } \psi_2 < x \leqslant \psi_3 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_n x + c + \beta_{n+1}(x - \Psi_{n+1})H(x - \Psi) + \zeta & \text{if } \psi_n < x \leqslant \psi_{n+1} \end{cases}$$

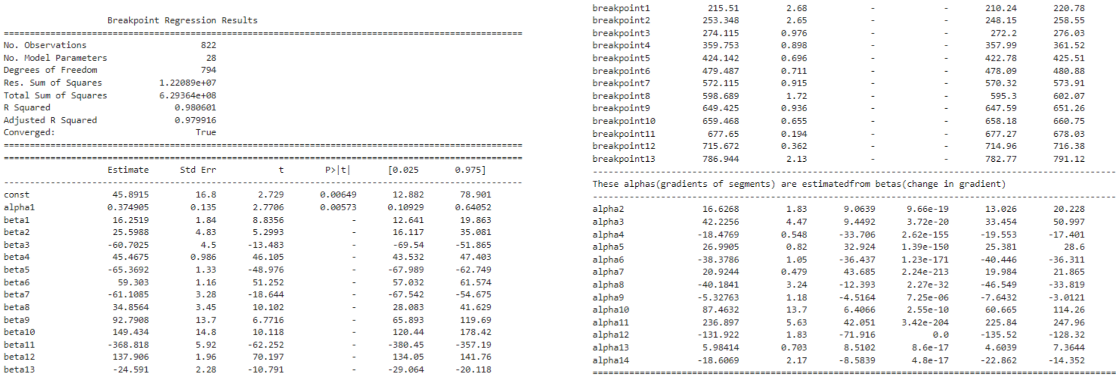
The model's general form with one breakpoint is:

$$y = \alpha x + c + \beta(x - \Psi)H(x - \Psi) + \zeta$$

With:

1. some data, x
2. α estimate the gradient of the i segment
3. c intercept of the i segment
4. β is the change in gradient from i to $i+1$ segments.(ie The the points at which the data's behaviour fully changes.)
5. ψ is the breakpoint position
6. ζ is a noise term
7. H is the Heaviside step function; 0 or 1

Because f(x) is now linear, we can use the statsmodels Python module to find a new breakpoint estimate, ψ_1 [40]. We repeat this process until the breakpoint estimate converges, at which point the method is terminated. If there are numerous breakpoints to consider, the same approach is used, with a multivariate Taylor expansion based on an initial guess for each breakpoint.



```
                    Breakpoint Regression Results
====================================================================================================
No. Observations                      822
No. Model Parameters                   28
Degrees of Freedom                    794
Res. Sum of Squares           1.22089e+07
Total Sum of Squares          6.29364e+08
R Squared                        0.980601
Adjusted R Squared               0.979916
Converged:                           True
====================================================================================================
====================================================================================================
                    Estimate      Std Err            t        P>|t|       [0.025       0.975]
----------------------------------------------------------------------------------------------------
const                45.8915         16.8        2.729      0.00649       12.882       78.901
alpha1              0.374905        0.135       2.7706      0.00573      0.10929      0.64052
beta1                16.2519         1.84       8.8356            -       12.641       19.863
beta2                25.5988         4.83       5.2993            -       16.117       35.081
beta3               -60.7025          4.5      -13.483            -       -69.54      -51.865
beta4                45.4675        0.986       46.105            -       43.532       47.403
beta5               -65.3692         1.33      -48.976            -      -67.989      -62.749
beta6                 59.303         1.16       51.252            -       57.032       61.574
beta7               -61.1085         3.28      -18.644            -      -67.542      -54.675
beta8                34.8564         3.45       10.102            -       28.083       41.629
beta9                92.7908         13.7       6.7716            -       65.893       119.69
beta10               149.434         14.8       10.118            -       120.44       178.42
beta11              -368.818         5.92      -62.252            -      -380.45      -357.19
beta12               137.906         1.96       70.197            -       134.05       141.76
beta13               -24.591         2.28      -10.791            -      -29.064      -20.118
```

(a)

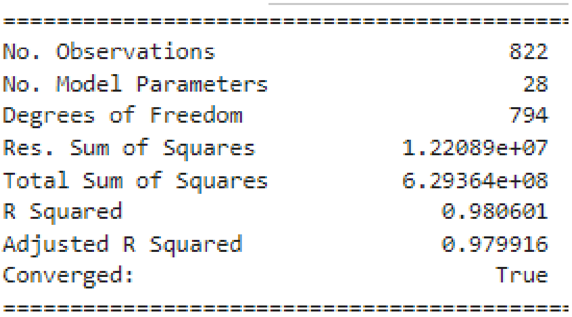
```
breakpoint1           215.51         2.68            -            -       210.24       220.78
breakpoint2          253.348         2.65            -            -       248.15       258.55
breakpoint3          274.115        0.976            -            -        272.2       276.03
breakpoint4          359.753        0.898            -            -       357.99       361.52
breakpoint5          424.142        0.696            -            -       422.78       425.51
breakpoint6          479.487        0.711            -            -       478.09       480.88
breakpoint7          572.115        0.915            -            -       570.32       573.91
breakpoint8          598.689         1.72            -            -        595.3       602.07
breakpoint9          649.425        0.936            -            -       647.59       651.26
breakpoint10         659.468        0.655            -            -       658.18       660.75
breakpoint11          677.65        0.194            -            -       677.27       678.03
breakpoint12         715.672        0.362            -            -       714.96       716.38
breakpoint13         786.944         2.13            -            -       782.77       791.12
----------------------------------------------------------------------------------------------------
These alphas(gradients of segments) are estimatedfrom betas(change in gradient)
----------------------------------------------------------------------------------------------------
alpha2               16.6268         1.83       9.0639     9.66e-19       13.026       20.228
alpha3               42.2256         4.47       9.4492     3.72e-20       33.454       50.997
alpha4              -18.4769        0.548      -33.706    2.62e-155      -19.553      -17.401
alpha5               26.9905         0.82       32.924    1.39e-150       25.381         28.6
alpha6              -38.3786         1.05      -36.437    1.23e-171      -40.446      -36.311
alpha7               20.9244        0.479       43.685    2.24e-213       19.984       21.865
alpha8              -40.1841         3.24      -12.393     2.27e-32      -46.549      -33.819
alpha9              -5.32763         1.18      -4.5164     7.25e-06      -7.6432      -3.0121
alpha10              87.4632         13.7       6.4066     2.55e-10       60.665       114.26
alpha11              236.897         5.63       42.051    3.42e-204       225.84       247.96
alpha12             -131.922         1.83      -71.916          0.0      -135.52      -128.32
alpha13              5.98414        0.703       8.5102      8.6e-17       4.6039       7.3644
alpha14             -18.6069         2.17      -8.5839      4.8e-17      -22.862      -14.352
====================================================================================================
```

(b)

Figure 6.2: Covid-19 infected cases Breakpoint Regression Results

6.4 Bayesian approach to linear modelling

The goal of Bayesian Linear Regression is to ascertain the posterior distribution for the model parameters rather than to identify the one "best" value of the model parameters.



```
==============================================
No. Observations                           822
No. Model Parameters                        28
Degrees of Freedom                         794
Res. Sum of Squares                1.22089e+07
Total Sum of Squares               6.29364e+08
R Squared                             0.980601
Adjusted R Squared                    0.979916
Converged:                                True
==============================================
```

Figure 6.3: Regression Results Table

In addition to the model parameters also coming from a distribution, the response is also generated from a probability distribution. The training inputs and outputs determine the posterior probability of the model parameters [41]:

$$P(\frac{\beta}{y}, X) = \frac{(P(\frac{\beta}{y}, X) * P(\frac{\beta}{X})}{P(\frac{y}{X})}$$

$$Posterior = \frac{Likelihood * Prior}{Normalization}$$

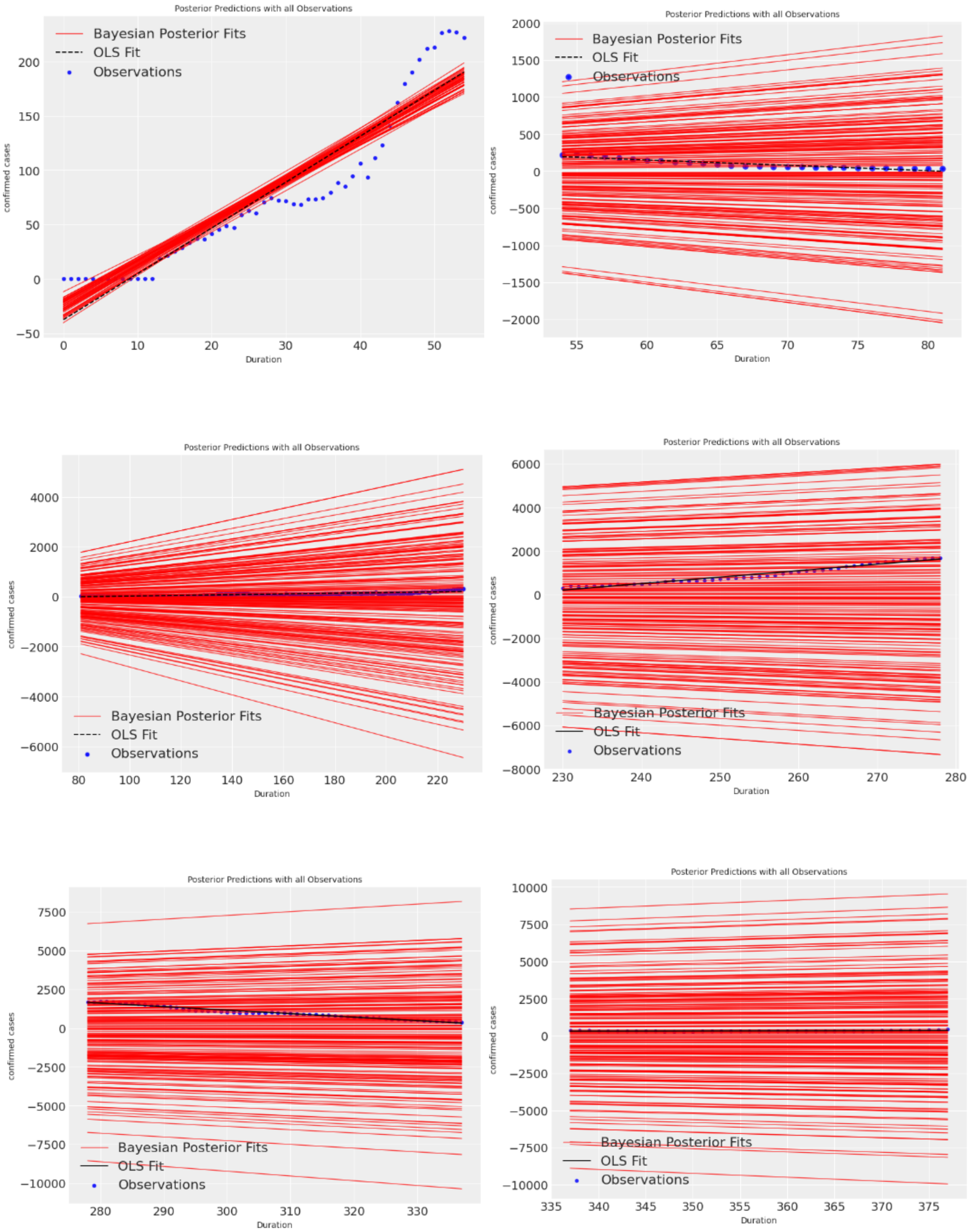
The model parameters' posterior probability distribution given the inputs and outputs is $P(\frac{\beta}{y}, X)$. This is equal to the likelihood of the data divided by a normalization constant, multiplied by the prior probability of the parameters. This is a straightforward formulation of the Bayes Theorem, which serves as the cornerstone of Bayesian inference.

6.5 Forecast for COVID-19 using Markov Chain Monte Carlo for Canada

The goal of the modelling is to estimate several scenarios for Covid-19 distribution in Canada. We start by inferring the parameters that best describe the observed condition, and then we use those parameters to predict future events. Monte Carlo importance sampling is used on the model parameters to infer a distribution of parameters that well describes the observed data for parameter estimation. For the forecast, we use parameter samples from this distribution to evolve the model equations [42].

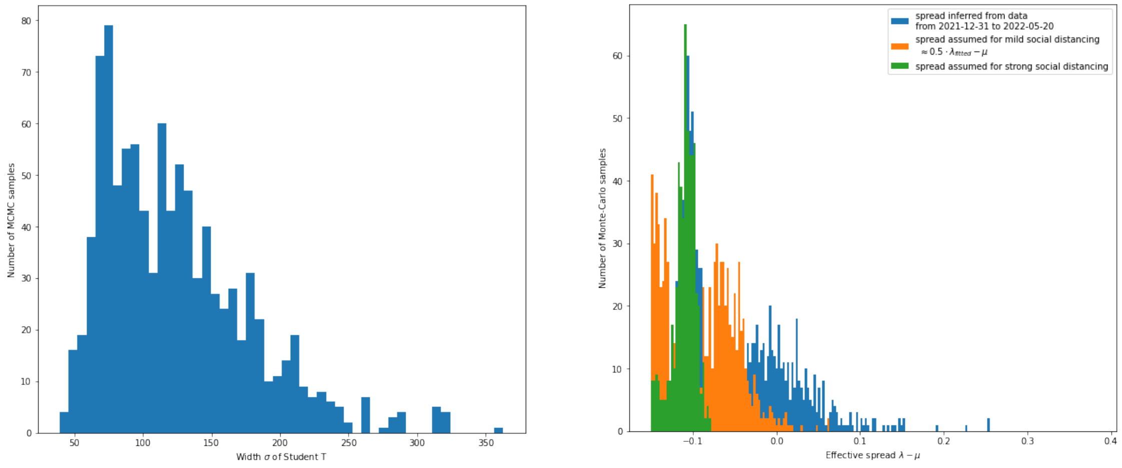
The data is insufficiently informative to fit all free parameters or to discover the underlying distribution experimentally. On the initial model rates, we establish the following information priors:

- The spreading rate is set to $\lambda \sim LogNormal(\log(0.4), 0.5)$, where 0.4 represents an estimate of 40% new infections every day.


Posterior Predictions with all Observations
Bayesian Posterior Fits
OLS Fit
Observations
confirmed cases
Duration
200
150
100
50
0
−50
0
10
20
30
40
50
Posterior Predictions with all Observations
Bayesian Posterior Fits
OLS Fit
Observations
confirmed cases
Duration
2000
1500
1000
500
0
−500
−1000
−1500
−2000
55
60
65
70
75
80
Posterior Predictions with all Observations
Bayesian Posterior Fits
OLS Fit
Observations
confirmed cases
Duration
4000
2000
0
−2000
−4000
−6000
80
100
120
140
160
180
200
220
Posterior Predictions with all Observations
Bayesian Posterior Fits
OLS Fit
Observations
confirmed cases
Duration
6000
4000
2000
0
−2000
−4000
−6000
−8000
230
240
250
260
270
280
Posterior Predictions with all Observations
Bayesian Posterior Fits
OLS Fit
Observations
confirmed cases
Duration
7500
5000
2500
0
−2500
−5000
−7500
−10000
280
290
300
310
320
330
Posterior Predictions with all Observations
Bayesian Posterior Fits
OLS Fit
Observations
confirmed cases
Duration
10000
7500
5000
2500
0
−2500
−5000
−7500
−10000
335
340
345
350
355
360
365
370
375

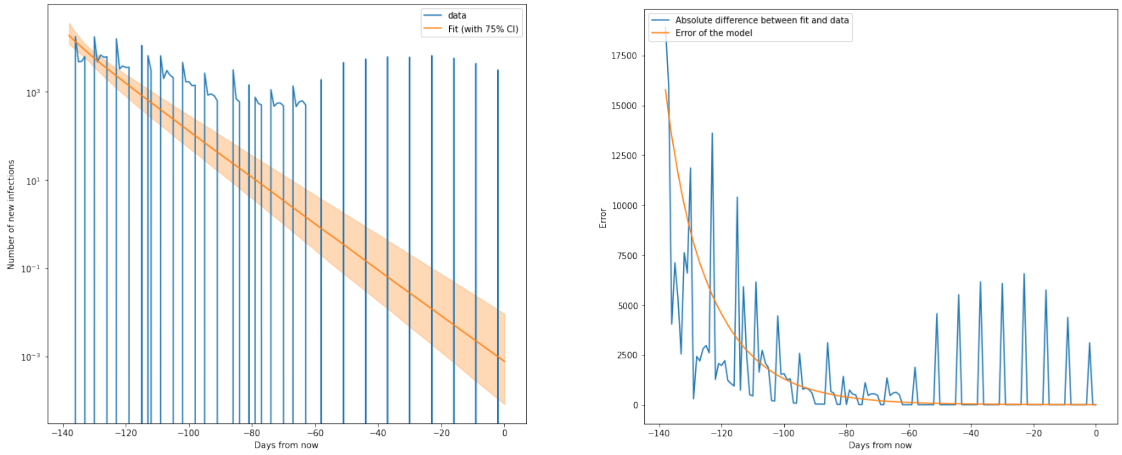
- The recovery rate is set at $\mu \sim LogNormal(\log(1/8), 0.2)$, which corresponds to an 8-day average recovery time.

Uninformative priors, in this case, the Half-Cauchy distribution, constrain the remaining model parameters. The MCMC sampler identifies the posterior distribution $p(\theta|I_{new})$ of model parameters θ that matches the real-world data. The effective spread ($\lambda - \mu$, which corresponds to the daily cases rate) derived from the data is plotted below as an example [42].



(a) (b)

Figure 6.4: Width σ of Student T and Effective spread $\lambda - \mu$



(a) (b)

Figure 6.5: Prediction and error of the model for infected cases

6.6 Bayesian Parameter Inference for Alberta's cases with pymc3

The purpose of Bayesian parameter inference is to estimate underlying parameter probability distributions from observable data.
First, we suppose that the data from is normal distribution, and use the Python statistics library to get the mean and The standard deviation 'stdev'
The Bayes graphical model for these data is shown in figure 6.6.

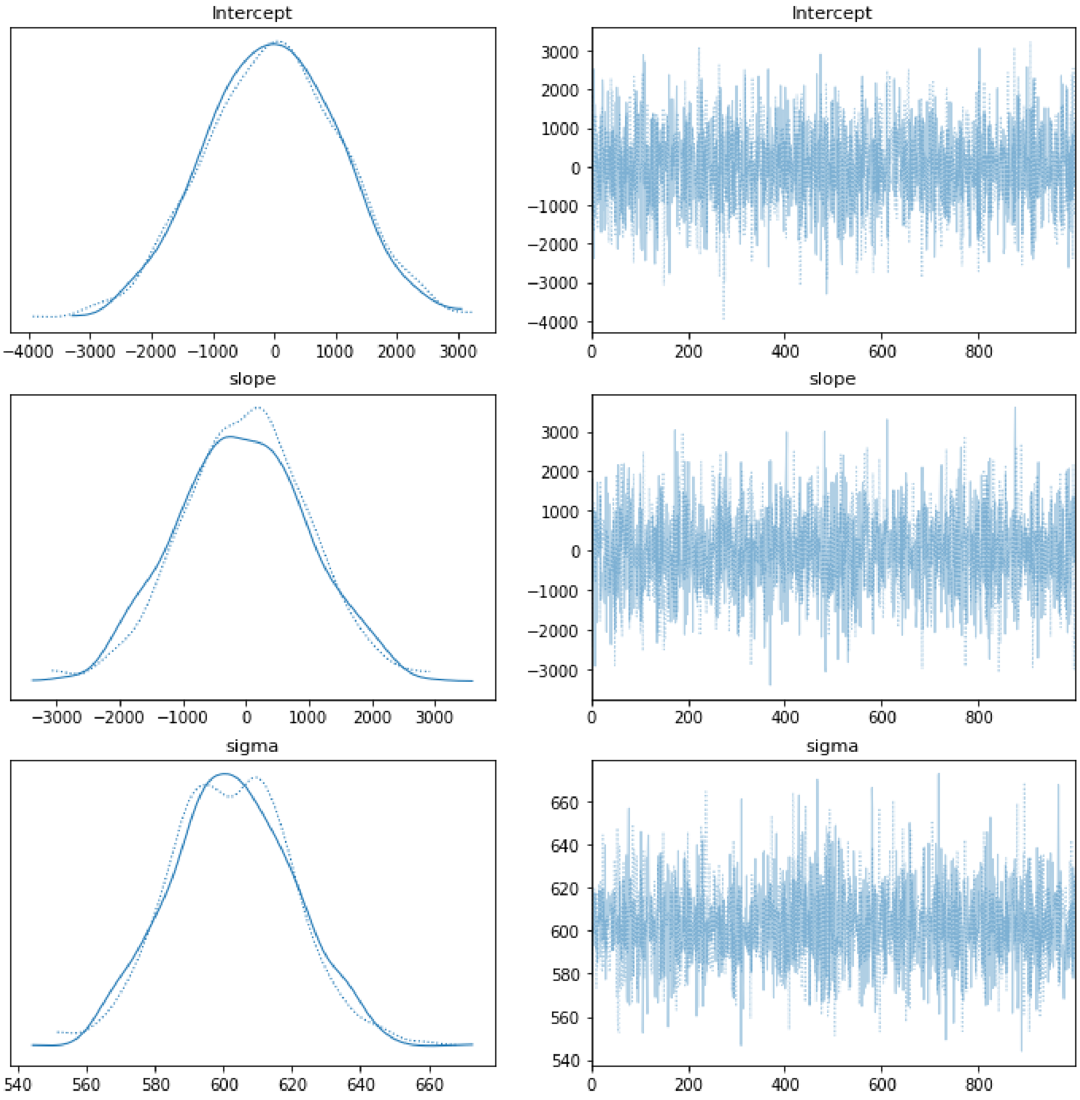


Figure 6.6: Bayes graphical model

Chapter 7

Unifying the Epidemiological and AI-based Modeling

7.1 Growth rate

To more accurately predict the spread of a disease or determine the basic reproduction number of the disease (R0), epidemiologists and public health experts measure the infection growth rates [43]. Growth rates have the advantage of being less prone to overfitting even though they lack key information that other statistics can provide. The growth rate can be evaluated using the following equation: $growth_rate = (\frac{present}{past})^{1/n} - 1$ where n = number of time periods.

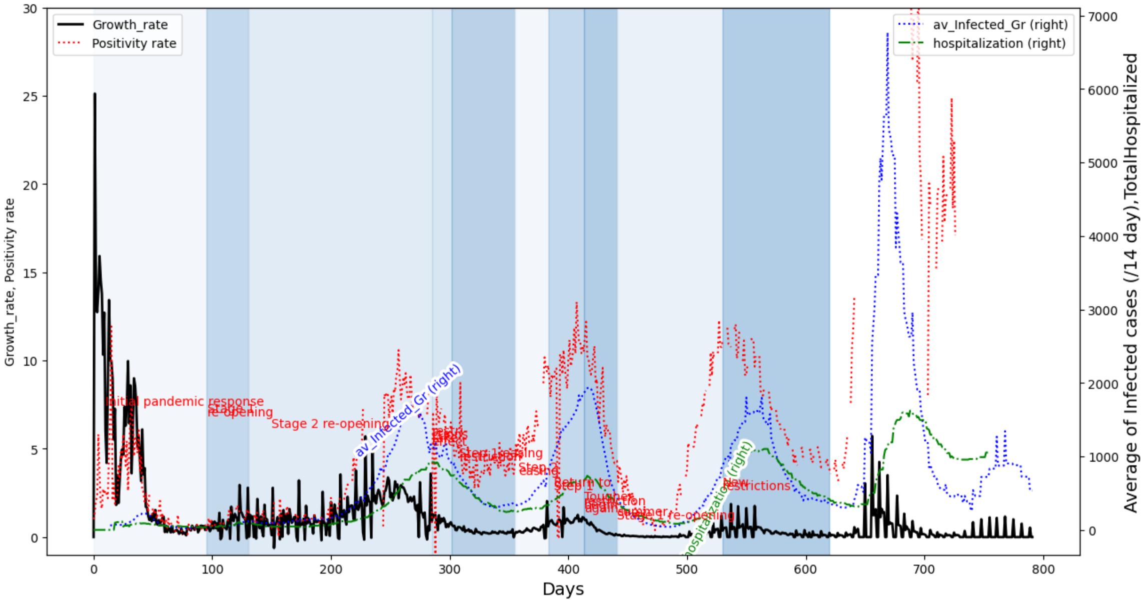


Figure 7.1: Alberta's daily cases Growth rate.

7.2 Positivity rate

The positivity rate can be expressed as (positive tests)/(total tests) x 100 percent or as the proportion of all coronavirus tests that are positive. The percent positive, also known as the "percent positive rate" or positivity rate, aids in the resolution of issues like the following for public health officials:

- Are we conducting sufficient testing given the number of infections?
- What is the rate of coronavirus transmission right now?

From figure 7.2 we can see that if there are too many positive tests or not enough tests overall, the % positives will be high. A greater percentage of positives denotes more transmission and the likelihood of a larger population of undiagnosed coronavirus carriers in the area.
From the above, it is clear that the % positive is an important metric since it shows us how prevalent the infection is in the region where testing is taking place and if testing levels are keeping up with the rate of disease transmission.

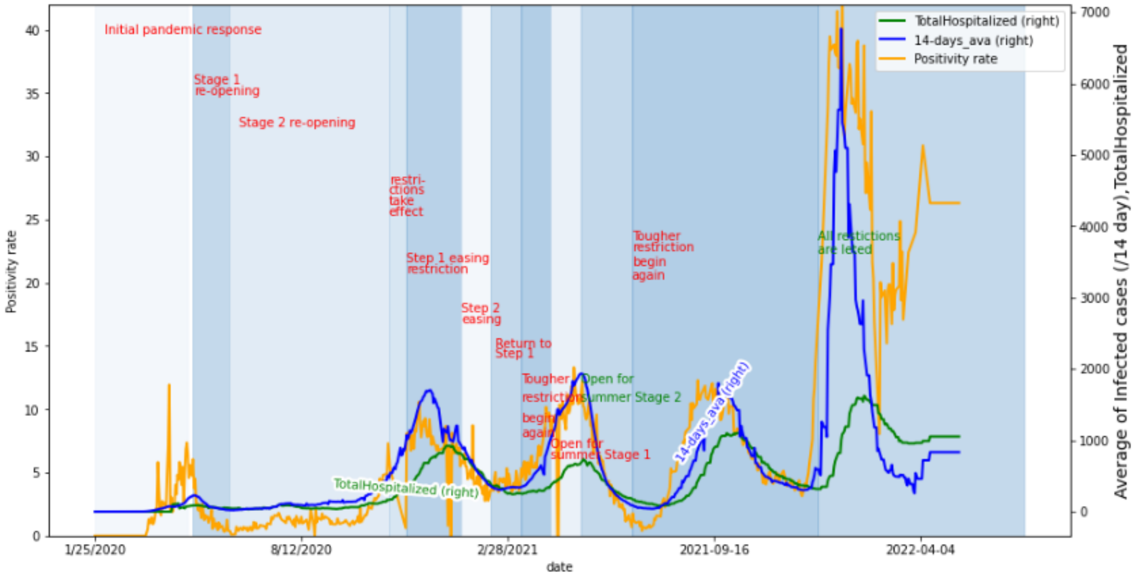


Figure 7.2: Covid-19 Infected cases, positivity rate, and hospitalized cases in Alberta

7.3 Transmission rate

As described in chapter 4, the transmission rate is the likelihood that an illness will spread among vulnerable individuals within a certain population. It is a crucial measure for showing how social interactions connect to the risk of transmission.

According to [44], nine instances were recorded in [45], with a rate of 35% (95 percent CI 27–44), depending on the type of contact that caused the illness.

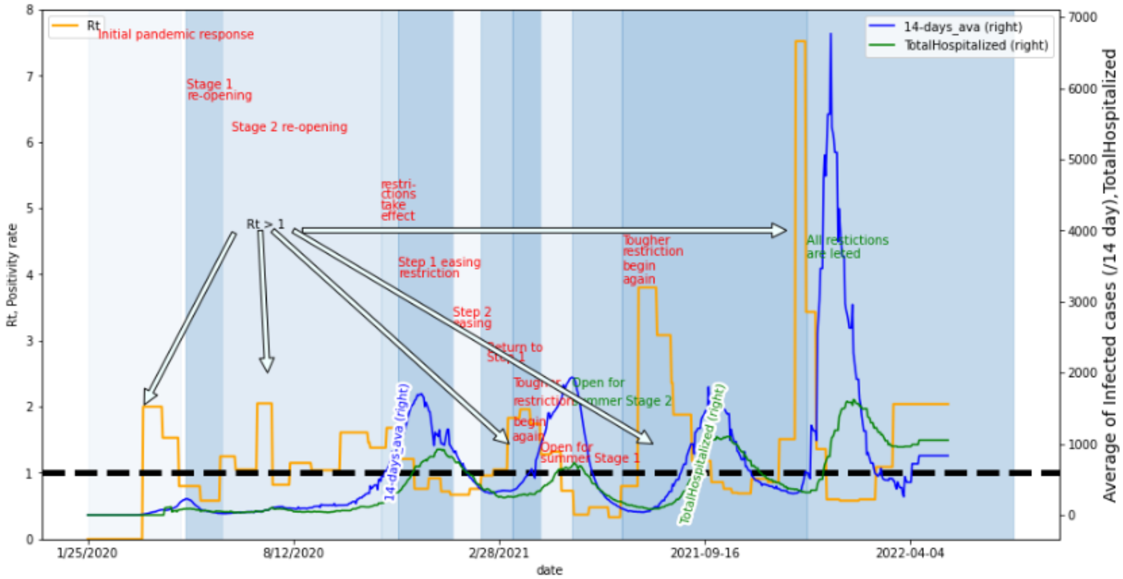


Figure 7.3: Covid-19 Infected cases, transmission rate, and hospitalized cases in Alberta

The SEIR (susceptible-exposed-infected-recovered) model, which has several generalisations, is the most well-known model in infectious disease epidemiology. To examine the strategic choices or efficacy of the mitigation measures, these models are used at the population level for the percentage of each state at a particular period [44]. Figure 7.3 serves as an example of the relationship between the number of infected cases and Alberta's transmission rate. We should anticipate an increase in the number of infected, ICU, and hospitalized cases when the transmission rate remains higher than 1.

We compute and graph the data with the transmission rate as we continue to analyze the data using ML. According to the decisions made by the government, every step is highlighted in figure 7.3. We can tell that the infection rate was higher than 1 for a considerable amount of time prior to the government taking action since the number of hospitalized patients is closely proportional to the number of infected cases. With ML, we can identify the issues and then modify or update the policy to prevent disasters.

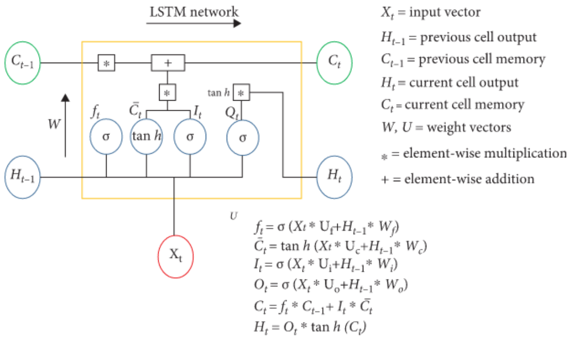
Chapter 8

Deep Learning Approach for COVID-19 Prediction.

In this chapter, we suggest an MTS-LSTM network that can simultaneously anticipate confirmed cases at the county level utilizing several time series and multiple variables.

8.1 LSTM

As illustrated in figure 8.3 [46], long short-term memory networks, or LSTMs, are employed in deep learning. Many recurrent neural networks (RNNs) can learn long-term dependencies, particularly in tasks involving sequence prediction. Except for singular data points like pictures, LSTM can analyze the full sequence of data and has feedback links. This has uses in machine translation and speech recognition, among others. A unique version of RNN called LSTM exhibits exceptional performance on a wide range of issues.



LSTM network
C_{t-1}
C_t
H_{t-1}
H_t
X_t
W
U
f_t
$\bar{C}_t$
I_t
Q_t
σ
tan h
σ
σ
tan h
*
+
*
*
X_t = input vector
H_{t-1} = previous cell output
C_{t-1} = previous cell memory
H_t = current cell output
C_t = current cell memory
W, U = weight vectors
* = element-wise multiplication
+ = element-wise addition
$f_t = \sigma\ (Xt * U_f + H_{t-1} * W_f)$
$\bar{C}_t = \tan h\ (Xt * U_c + H_{t-1} * W_c)$
$I_t = \sigma\ (X_t * U_i + H_{t-1} * W_i)$
$O_t = \sigma\ (X_t * U_o + H_{t-1} * W_o)$
$C_t = f_t * C_{t-1} + I_t * \bar{C}_t$
$H_t = O_t * \tan h\ (C_t)$

Figure 8.1: The overall structure of the LSTM model.

8.2 Step-by-step LSM walkthrough

Predict the next infected case of Covid-19 using data from past cases. The gender of the current subject may be included in the cell state. [47].

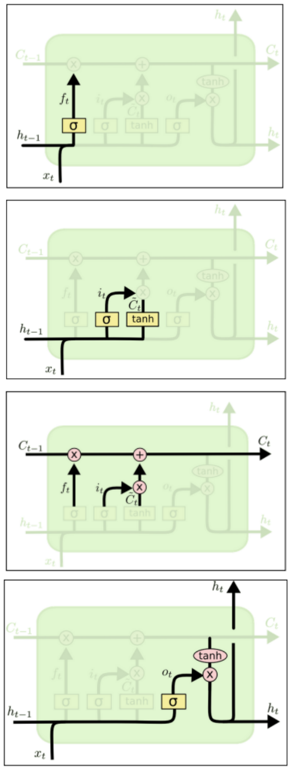


Figure 8.2: test

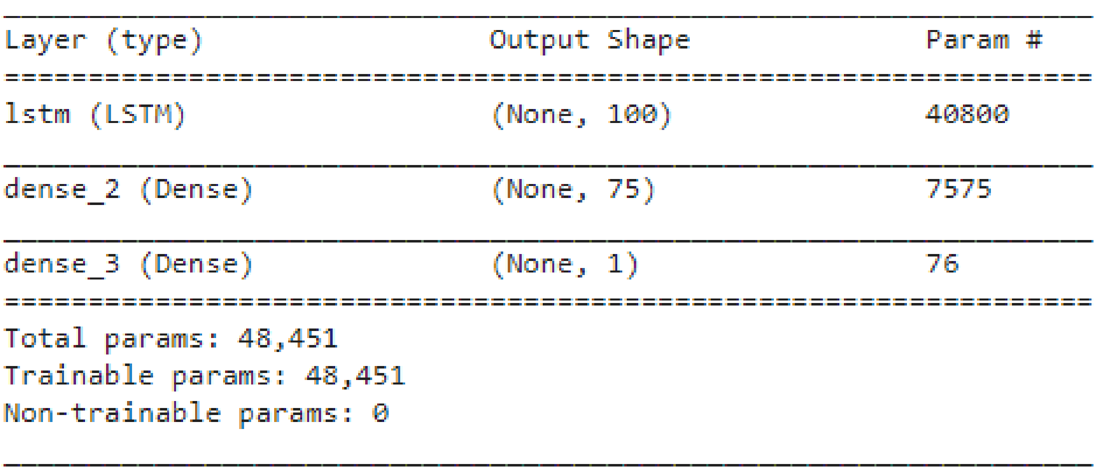
1. The first step is to discard cell state information.

2. The next step is to choose the new data that will be kept in the cell state.

3. Next Update old cell state Ct-1 to new cell state Ct at this moment.

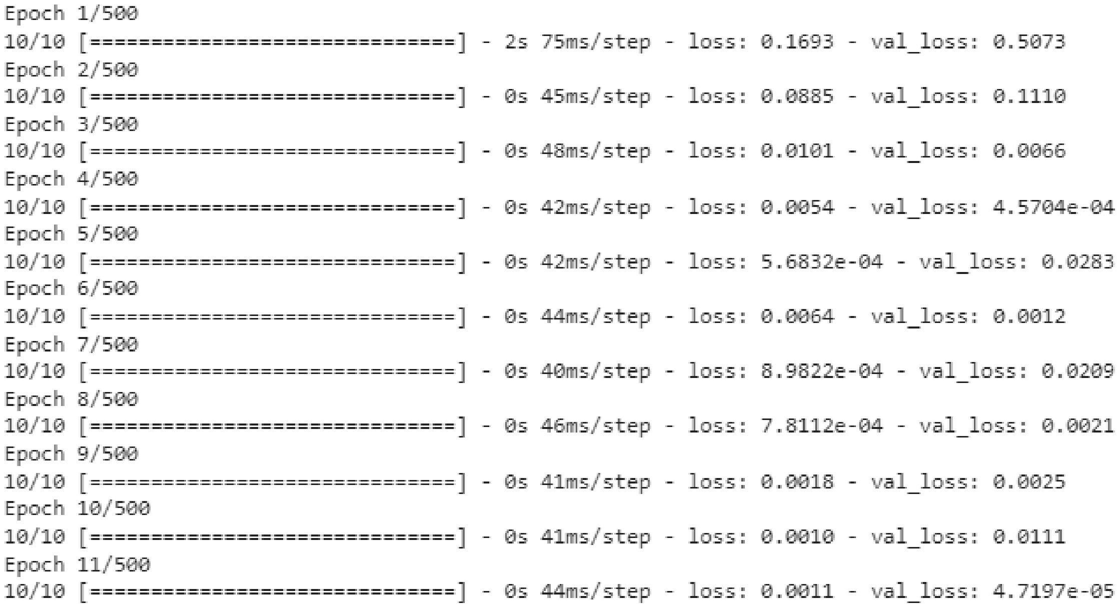
4. Finally, we decide on our output.

The following results are obtained when the aforementioned procedures are used with a group of infected cases:

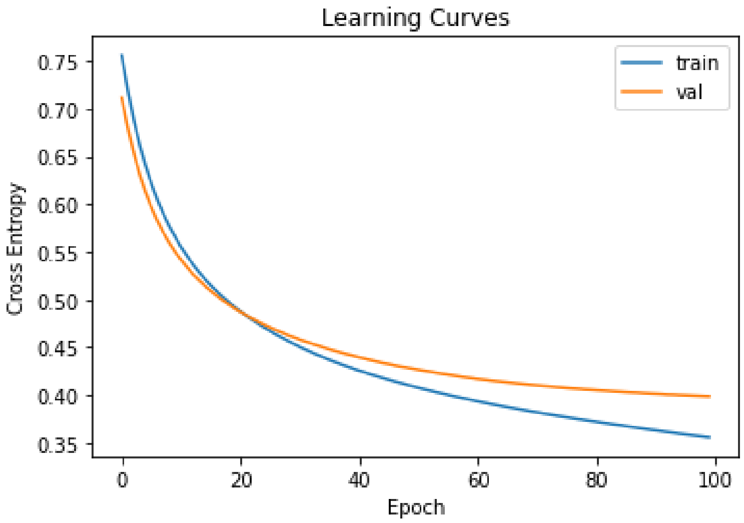


```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 100)               40800

_________________________________________________________________
dense_2 (Dense)              (None, 75)                7575

_________________________________________________________________
dense_3 (Dense)              (None, 1)                 76
=================================================================
Total params: 48,451
Trainable params: 48,451
Non-trainable params: 0

_________________________________________________________________
```

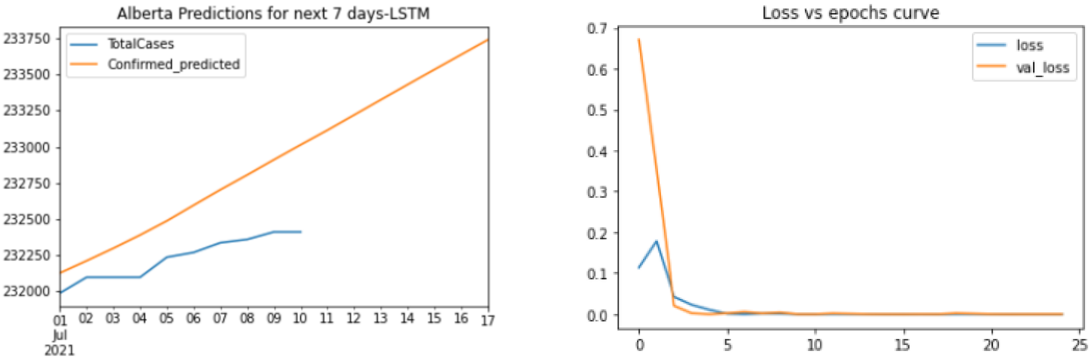
Figure 8.3: A recurrent neural network (LSTM) model with two hidden layers containing 150+75 nodes.



```
Epoch 1/500
10/10 [==============================] - 2s 75ms/step - loss: 0.1693 - val_loss: 0.5073
Epoch 2/500
10/10 [==============================] - 0s 45ms/step - loss: 0.0885 - val_loss: 0.1110
Epoch 3/500
10/10 [==============================] - 0s 48ms/step - loss: 0.0101 - val_loss: 0.0066
Epoch 4/500
10/10 [==============================] - 0s 42ms/step - loss: 0.0054 - val_loss: 4.5704e-04
Epoch 5/500
10/10 [==============================] - 0s 42ms/step - loss: 5.6832e-04 - val_loss: 0.0283
Epoch 6/500
10/10 [==============================] - 0s 44ms/step - loss: 0.0064 - val_loss: 0.0012
Epoch 7/500
10/10 [==============================] - 0s 40ms/step - loss: 8.9822e-04 - val_loss: 0.0209
Epoch 8/500
10/10 [==============================] - 0s 46ms/step - loss: 7.8112e-04 - val_loss: 0.0021
Epoch 9/500
10/10 [==============================] - 0s 41ms/step - loss: 0.0018 - val_loss: 0.0025
Epoch 10/500
10/10 [==============================] - 0s 41ms/step - loss: 0.0010 - val_loss: 0.0111
Epoch 11/500
10/10 [==============================] - 0s 44ms/step - loss: 0.0011 - val_loss: 4.7197e-05
```



$$MSE : 3889857024.000, \qquad RMSE : 62368.718$$



(a) (b)

Figure 8.4: Prediction and error of the model for infected cases

Chapter 9

Time Series Analysis for COVID-19 Disease Prediction.

A time series is a collection of data points that are measured sequentially, usually spanning time intervals. Time series analysis refers to techniques for deriving useful statistics and other aspects of time series data through analysis.
Depending on whether the current value of the series is a linear or non-linear consequence of earlier observations, a time series model is referred to be linear or non-linear.

- Components of a Time Series

$$Y(t) = T(t) + S(t) + C(t) + I(t)$$

$$trend(t) + seasonal(t) + cyclical(t) + irregular_{unpredictableinfluences}(t)$$

$$Example : RandomWalk$$

9.1 Cumulative cases as linear regression modeling

The foundation of statistical modelling is time series linear regression, and that is our starting point in this section. It is well known that the simplest model to represent the regression function as a linear combination of predictors is linear regression. The model parameters are simple to grasp due to the linear shape. Additionally, mathematically elegant linear model theories are widely known. Furthermore, a lot of contemporary modelling tools are built on the foundation of linear regression. For instance, linear regression frequently offers a good approximation to the underlying regression function, especially when the sample size is small or the signal is very faint [48]. Figure 9.1 illustrates the linear regression model of the cumulative infected cases in Alberta.

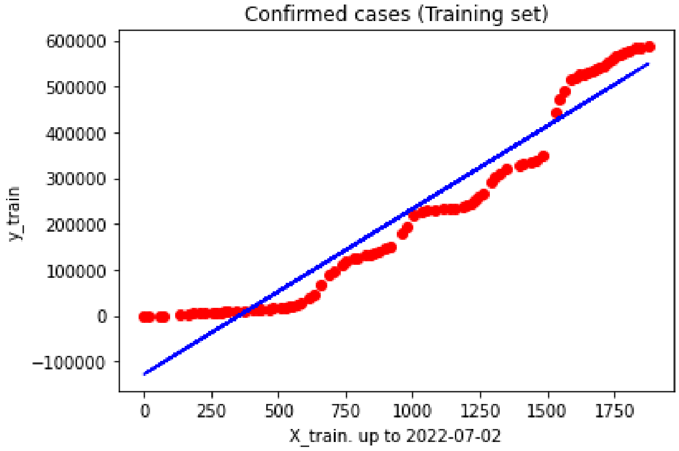


Figure 9.1: Alberta linear regression

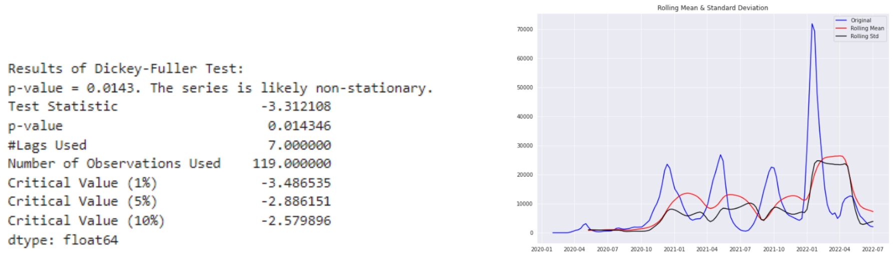
9.2 Augmented Dickey–Fuller test (ADF)

A Dickey-Fuller test is a unit root test that examines the null hypothesis that $\alpha = 1$ in the model equation below [α is the coefficient of the first lag on Y. The interval between the two-time series you are correlating is known as the lag time. The lag time would be 1 if the autocorrelation of the data sets $(0, 1), (1, 2)...(n - 1, n)$ were taken apart].

$$y_t = c + \beta t + \alpha y_{t-1} + \phi Y_{t-1} + e_t$$

with y_{t-1} = lag 1 of time series.
ϕY_{t-1} = first difference of the series at time (t-1) [49].

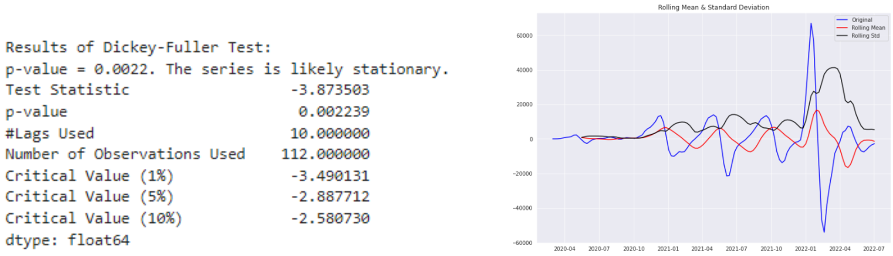
Figure 9.2 shows the results of a Python code that calculates the mean and standard deviation of the series and runs the enhanced Covid-19 infected cases test. The pvale is returned. The series is more stationary when the pvalue is small.
Figure 9.3 represent the same data after differencing (it is a technique for changing a time series dataset. It can be used to get rid of the series' so-called temporal reliance on time), and we can see that the p-value went from 0.014346 to 0.002239.



```
Results of Dickey-Fuller Test:
p-value = 0.0143. The series is likely non-stationary.
Test Statistic                  -3.312108
p-value                          0.014346
#Lags Used                       7.000000
Number of Observations Used    119.000000
Critical Value (1%)             -3.486535
Critical Value (5%)             -2.886151
Critical Value (10%)            -2.579896
dtype: float64
```

(a) (b)

Figure 9.2: Results of Dickey-Fuller Test



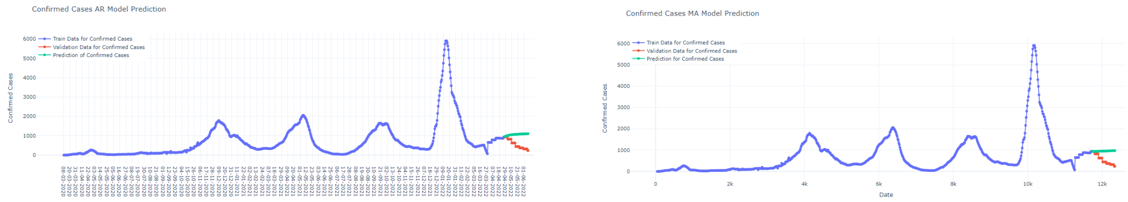
```
Results of Dickey-Fuller Test:
p-value = 0.0022. The series is likely stationary.
Test Statistic                  -3.873503
p-value                          0.002239
#Lags Used                      10.000000
Number of Observations Used    112.000000
Critical Value (1%)             -3.490131
Critical Value (5%)             -2.887712
Critical Value (10%)            -2.580730
dtype: float64
```

(a) (b)

Figure 9.3: Results of Dickey-Fuller after differencing

9.3 ARIMA prediction

Auto-Regressive Integrated Moving Average is referred to as ARIMA. in particular, **AR** Autoregression. Figure 9.4(a) is a model that takes into account the dependency between an observation and a certain number of lag observations.
I combined. using differencing to make the time series stable by using differentiating raw observations.
MA Average movement. Figure 9.4(b) is a model that takes advantage of the relationship between a lagged observation and a residual error from a moving average model.



(a) (b)

Figure 9.4: AR and MA output

Figure 9.5 illustrates the whole ARIMA output for confirmed cases prediction.

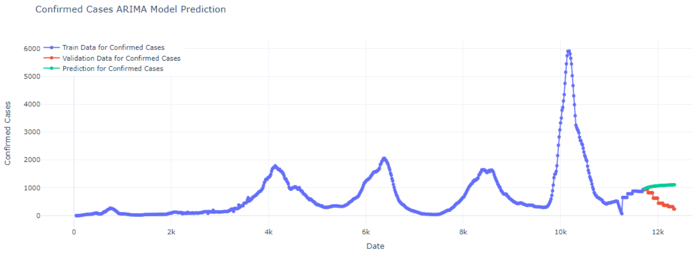


Figure 9.5: Alberta ARIMA Forecasting

9.4 SARIMA prediction

The seasonal autoregressive integrated moving average (SARIMA) model was used to anticipate the incidence of dengue using R software. Using data gathered between January 2020 and October 2022, we validated the model after fitting it using the daily infected cases reported in Alberta. RESULTS: The model with the greatest data fit was SARIMA, as shown in Figure 9.6. Figures for 2022 are anticipated to be quite similar to the actual ones.

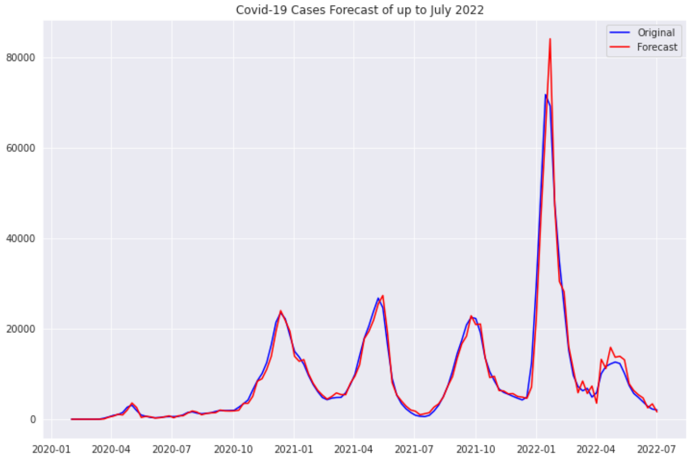
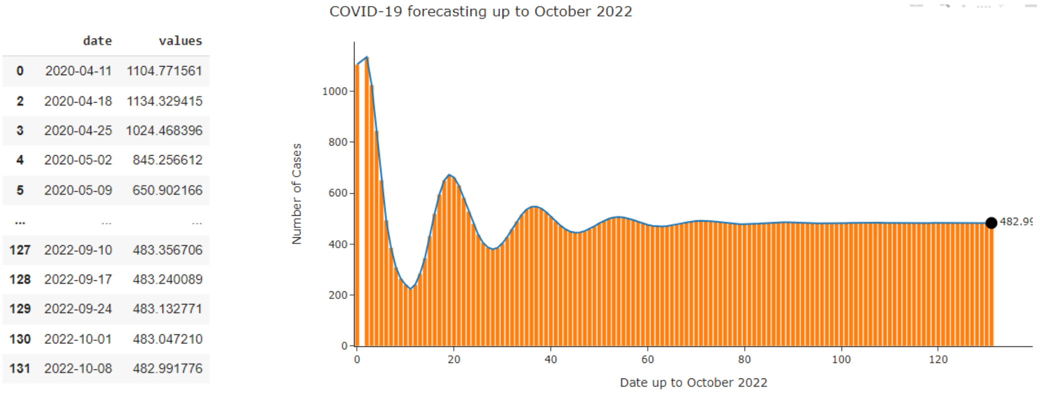


Figure 9.6: Alberta Forecasting



	date	values
0	2020-04-11	1104.771561
2	2020-04-18	1134.329415
3	2020-04-25	1024.468396
4	2020-05-02	845.256612
5	2020-05-09	650.902166
...	...	...
127	2022-09-10	483.356706
128	2022-09-17	483.240089
129	2022-09-24	483.132771
130	2022-10-01	483.047210
131	2022-10-08	482.991776

(a)

(b)

Figure 9.7: Prediction and error of the model for infected cases

9.4.1 SARIMA forecasting

For the forecasting of the next two years, we get the output shown in figure 9.8

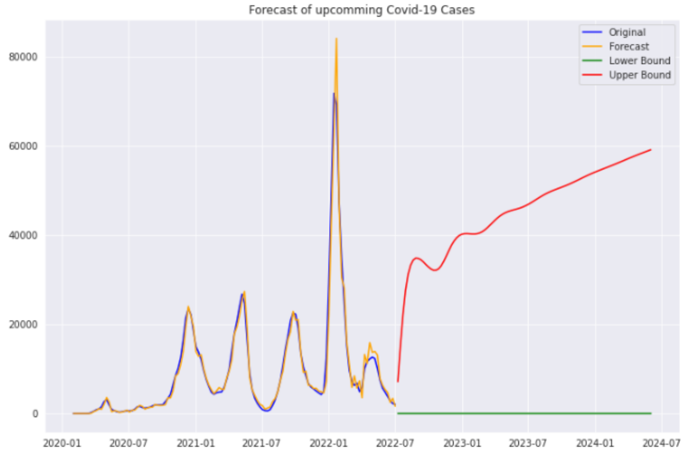


Figure 9.8: SARIMA forcating

9.4.2 Interpretation of ACF and PACF plots for Identifying ARIMA Model

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are crucial in time series analysis for supplying model orders like p for SARIMA to choose the optimal model for predicting. ACF between time series and a lagged version of itself. The relationship between observations made at different times. The autocorrelation function begins with a lag of zero, which is the correlation of the time series with itself, resulting in a correlation of one.

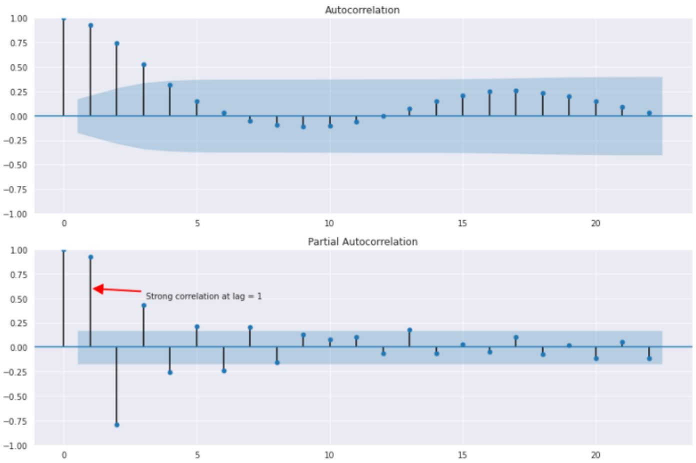


Figure 9.9: AFC and PACF

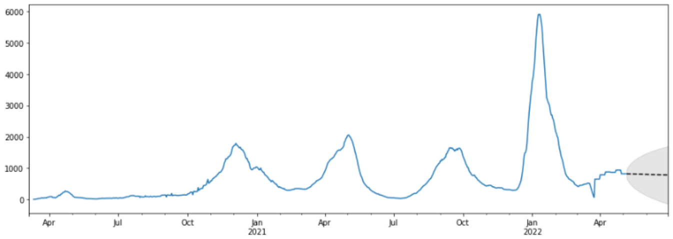


Figure 9.10: Alberta Forecasting

9.5 Holt-Winters forecasting (HW)

The Holt-Winters technique is a statistical forecasting approach for univariate time series. Forecasting is making predictions about future performance based on previous and recent data. Forecasting seasonal time series is frequently done using Holt-Winters exponential smoothing. Because they model the level, trend, and seasonality of a time series, the Winters approach and Fourier series analysis are flexible techniques. The following equations outline the additive Holt-Winters approach [50].

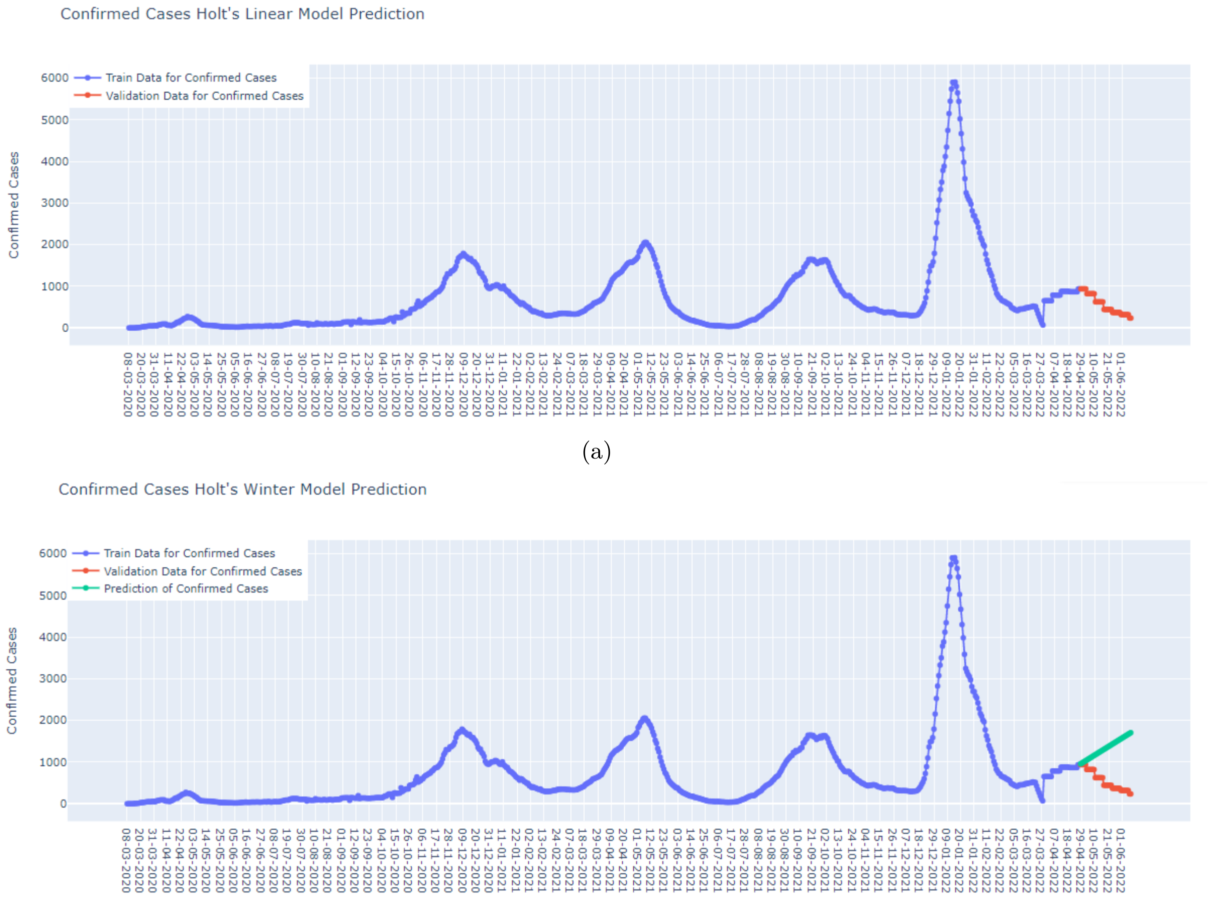
$$\beta_0 + \beta_1 t + sn_{t+\epsilon_t}$$

Estimate the level at time T as:

$$l_t = \alpha(y_t - sn_{T-L}) + (1-\alpha)(l_{T-1} + b_{T-1})$$

Estimate of the growth rate (or trend) at time T:

$$b_T = \gamma(l_T - l_{T-1}) + (1-\gamma)b_{T-1}; \quad 0 \leq \alpha; \gamma \leq 1$$



(a)

(b)

Figure 9.11: HOLT's daily case predicting model

Chapter 10

Control Chart and Filtering for COVID-19 Disease Prediction.

10.1 Exponentially weighted moving average (EWMA)

The exponentially weighted moving average (EWMA) is frequently applied to a time-ordered sequence of random variables. By applying weights that decrease geometrically with the age of the data, it calculates a weighted average of the sequence. The EWMA is defined by [51]:
Consider the n × 1 random vector x given by $x = [x_1, x_2, ..., x_n]$, the linear transformation:

$$z = Cx + z_0 b,$$

where: $C = \begin{bmatrix} \lambda & 0 & 0 & ... & 0 \\ \lambda(1-\lambda) & \lambda & 20 & ... & 0 \\ \lambda(1-\lambda)^2 & \lambda(1-\lambda) & \lambda & ... & 0 \\ ... & ... & .... & ... & 0 \\ \lambda(1-\lambda)^{n-1} & \lambda(1-\lambda)^{n-2} & ... & ... & \lambda \end{bmatrix}$;

b is a known n × 1 vector having the form:

$$b = ((1-\lambda)^1 (1-\lambda)^2 (1-\lambda)^3 ....(1-\lambda)^n);$$

z_0 is an initial (scalar) value that represents the EWMA's starting value.
The parameter $\lambda (0 < \lambda \leq 1)$ is known as the smoothing coefficient, and its value is frequently chosen in practice based on how quickly the process means changes [51].

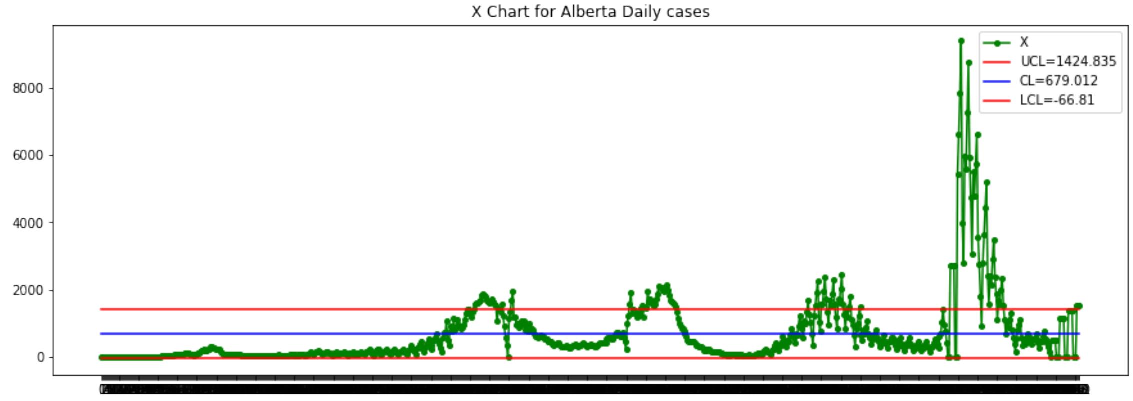
10.2 Steps in constructing a control chart

Let's utilize a straightforward equation of a control chart with Alberta's Covid-19 daily cases to make things easier to grasp. Figure **??** shows UCL as the upper control limit and LCL as the lower control limit.

$$\overline{\mathrm{x}} = \frac{\Sigma x}{k} \quad (CL)$$

$$UCL = \overline{\mathrm{x}} + 3\sqrt{\overline{\mathrm{x}}}$$

$$LCL = \overline{\mathrm{x}} - 3\sqrt{\overline{\mathrm{x}}}$$



10.3 EWMA Covid-19 application

To illustrate the EWMA, Figure 10.1 shows a plot of Alberta's Covid-19 infected cases using pyspc Python library. 3,6, and 12 Span specify decay in terms of span.

Figure 10.2 can help us better understand when and where the Covid-19 situation is critical. We can argue that there is a difficulty in controlling the issue when the number of infected cases or the number of hospitalized cases is outside of the green region, and some action should be performed.

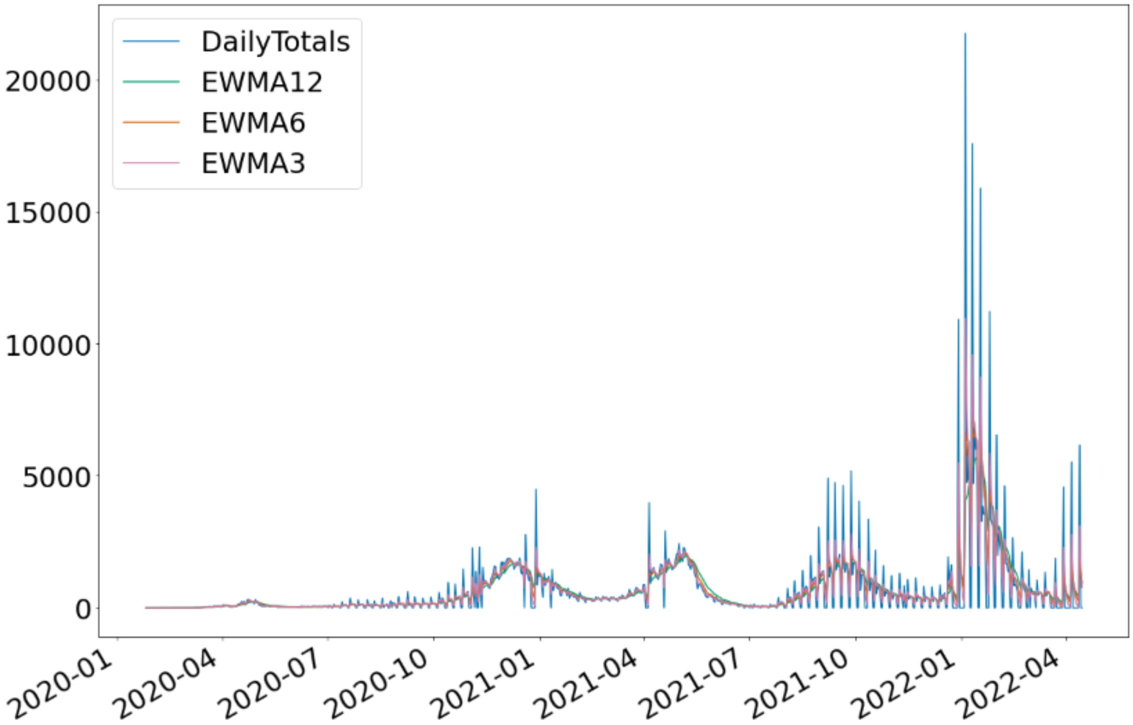


Figure 10.1: Alberta's daily cases EWMA

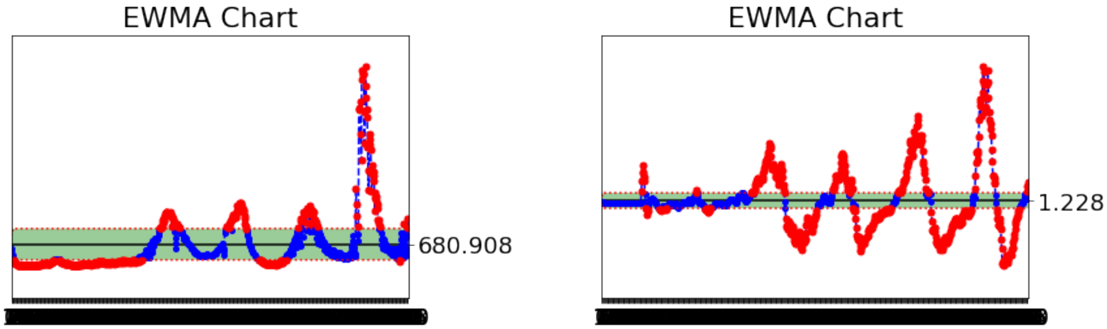


Figure 10.2: Alberta's daily infected & hospitalized cases.

10.4 Finding the EWMA of Alberta's transmission rate

Using the EWMA of the transmission rate, we can additionally keep track of the Covid-19 cases. The management of Covid-19 judgments based solely on Rt<1 and Rt>1 can be challenging but suffused green margins as chowed in figure 10.3 allow us to be more flexible in our decision-making.

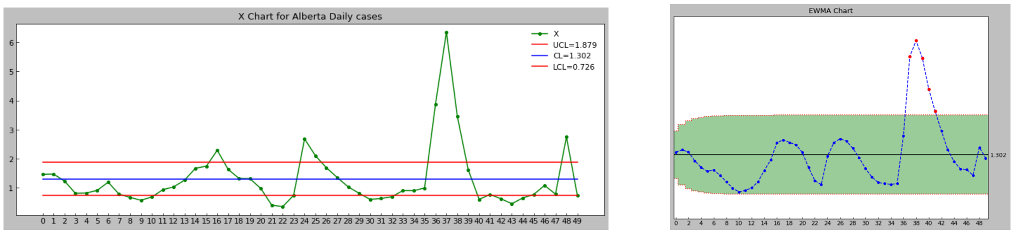


Figure 10.3: EWMA of transmission rate

$$UCL = 1.879 \qquad CL = 1.3 \quad LCL = 0.726$$

Chapter 11

Conclusion and Future Works.

11.1 Goal

This work's primary goal is to provide a summary of earlier research and its COVID-19 applications. The illustration and identification of the COVID-19 epidemic infected cases using deep learning and machine learning methods are covered in detail in this publication. With the use of machine learning and deep learning techniques, we did our best to steer clear of duplicate concepts with content that was pertinent to COVID-19. Generally speaking, machine learning and deep learning methodologies are used to evaluate and interpret the COVID-19 summary findings.

11.2 Achievement

We used Python to apply practically all supervised learning algorithms, and we saw how one ML result might be quite helpful in directing us to choose the best choice. We wholeheartedly concur that some applications are difficult to use and comprehend, but the positive results will be worth it.

11.3 Future Works

Although some of us, including myself, believe that COVID-19 is over, this work will always be relevant and useful for other infectious diseases. We gained a lot of knowledge from Covid-19, and we are now documenting our learned lessons for the project that will begin once this one is over.

Bibliography

[1] S. Mohan, A. Abugabah, S. Kumar Singh, A. kashif Bashir, and L. Sanzogni, “An approach to forecast impact of covid-19 using supervised machine learning model,” *Software: Practice and Experience*, vol. 52, no. 4, pp. 824–840, 2022.

[2] medrxiv.org, *Health topics/coronavirus. 2021.* 2022. [Online]. Available: `https://covid19.who.int`.

[3] *World health organization. who timeline – covid-19. geneva (ch): Who; april 27, 2020.* 2022. [Online]. Available: `https://www.who.int/news-room/detail/08-04-2020-who-timeline---covid-19`.

[4] *Modelling scenarios of the epidemic of covid-19 in canada.* 2022. [Online]. Available: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7343050/`.

[5] L. Yang *et al.*, “Covid-19: Immunopathogenesis and immunotherapeutics,” *Signal transduction and targeted therapy*, vol. 5, no. 1, pp. 1–8, 2020.

[6] 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. 109 864, 2020.

[7] M. Masum, M. Masud, M. I. Adnan, H. Shahriar, and S. Kim, “Comparative study of a mathematical epidemic model, statistical modeling, and deep learning for covid-19 forecasting and management,” *Socio-Economic Planning Sciences*, vol. 80, p. 101 249, 2022.

[8] Y. Xiong, Y. Ma, L. Ruan, D. Li, C. Lu, and L. Huang, “Comparing different machine learning techniques for predicting covid-19 severity,” *Infectious diseases of poverty*, vol. 11, no. 1, pp. 1–9, 2022.

[9] D. Tiwari, B. S. Bhati, F. Al-Turjman, and B. Nagpal, “Pandemic coronavirus disease (covid-19): World effects analysis and prediction using machine-learning techniques,” *Expert Systems*, vol. 39, no. 3, e12714, 2022.

[10] R. Chandra, A. Jain, and D. Singh Chauhan, “Deep learning via lstm models for covid-19 infection forecasting in india,” *PloS one*, vol. 17, no. 1, e0262708, 2022.

[11] H. Abbasimehr, R. Paki, and A. Bahrini, “A novel approach based on combining deep learning models with statistical methods for covid-19 time series forecasting,” *Neural Computing and Applications*, vol. 34, no. 4, pp. 3135–3149, 2022.

[12] N. Hu *et al.*, “The impact of the covid-19 pandemic on paediatric health service use within one year after the first pandemic outbreak in new south wales australia–a time series analysis,” *The Lancet Regional Health-Western Pacific*, vol. 19, p. 100 311, 2022.

[13] L. Xie, “The analysis and forecasting covid-19 cases in the united states using bayesian structural time series models,” *Biostatistics & Epidemiology*, vol. 6, no. 1, pp. 1–15, 2022.

[14] J. Chen, J. J. Song, and J. D. Stamey, “A bayesian hierarchical spatial model to correct for misreporting in count data: Application to state-level covid-19 data in the united states,” *International Journal of Environmental Research and Public Health*, vol. 19, no. 6, p. 3327, 2022.

[15] N. Kianfar, M. S. Mesgari, A. Mollalo, and M. Kaveh, “Spatio-temporal modeling of covid-19 prevalence and mortality using artificial neural network algorithms,” *Spatial and Spatio-temporal Epidemiology*, vol. 40, p. 100 471, 2022.

[16] Á. Briz-Redón, A. Iftimi, J. F. Correcher, J. De Andrés, M. Lozano, and C. Romero-Garcıa, “A comparison of multiple neighborhood matrix specifications for spatio-temporal model fitting: A case study on covid-19 data,” *Stochastic Environmental Research and Risk Assessment*, vol. 36, no. 1, pp. 271–282, 2022.

[17] M. Jahja, A. Chin, and R. J. Tibshirani, “Real-time estimation of covid-19 infections: Deconvolution and sensor fusion,” *Statistical Science*, vol. 37, no. 2, pp. 207–228, 2022.

[18] J. Y. Choi, “Covid-19 in south korea,” *Postgraduate medical journal*, vol. 96, no. 1137, pp. 399–402, 2020.

[19] F. C. Stefano SCARPETTA Mark PEARSON, *Access to covid-19 vaccines: Global approaches in a global crisis*, 2021. [Online]. Available: `https://www.oecd.org/coronavirus/policy-responses/access-to-covid-19-vaccines-global-approaches-in-a-global-crisis-c6a18370/#boxsection-d1e30`.

[20] M. Eissa and E. Rashed, “Descriptive analysis of coronavirus disease cases based on geographical distribution in canadian provinces/territories: Statistical investigation into epidemiological pattern,” *Academia Letters*, p. 2, 2022.

[21] A. M. · , *Alberta acted like the pandemic was over. now it’s a cautionary tale for canada*, 2021. [Online]. Available: `https://www.cbc.ca/news/health/alberta-fourth-wave-surge-hospitals-icu-covid-19-1.6197263`.

[22] M. Garner, S. Hamilton, *et al.*, “Principles of epidemiological modelling,” *Revue Scientifique et Technique-OIE*, vol. 30, no. 2, p. 407, 2011.

[23] R. N. Thompson, “Epidemiological models are important tools for guiding covid-19 interventions,” *BMC medicine*, vol. 18, no. 1, pp. 1–4, 2020.

[24] K. Sasaki, *Covid-19 dynamics with sir model*, 2022. [Online]. Available: `https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html`.

[25] S. Ho, *Coronavirus stats worldwide: Compare canada and other key nations*, May 31, 2022. [Online]. Available: `https://www.ctvnews.ca/health/coronavirus/coronavirus-stats-worldwide-compare-canada-and-other-key-nations-1.4881500`.

[26] L. Xue, S. Jing, and H. Wang, “Evaluating the impacts of non-pharmaceutical interventions on the transmission dynamics of covid-19 in canada based on mobile network,” *PloS one*, vol. 16, no. 12, e0261424, 2021.

[27] P. Ravani *et al.*, "Covid-19 screening of asymptomatic patients admitted through emergency departments in alberta: A prospective quality-improvement study," *Canadian Medical Association Open Access Journal*, vol. 8, no. 4, E887–E894, 2020.

[28] J. G. Ajay Agrawal and A. Goldfarb, *How to win with machine learning*, Sep, 2020. [Online]. Available: `https://hbr.org/2020/09/how-to-win-with-machine-learning`.

[29] J. Le, *The top 10 machine learning algorithms every beginner should know*, March 16, 2022. [Online]. Available: `https://builtin.com/data-science/tour-top-10-algorithms-machine-learning-newbies`.

[30] N. J. Nilsson, *Introduction to machine learning*, November 3, 1998. [Online]. Available: `https://ai.stanford.edu/~nilsson/MLBOOK.pdf`.

[31] N. S. Chauhan, *Decision tree algorithm, explaine*, February 9, 2022. [Online]. Available: `https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html#:~:text=Decision%5C%20Trees%5C%20follow%5C%20Sum%5C%20of,form%5C%20a%5C%20disjunction%5C%20(sum).`.

[32] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geology Reviews*, vol. 71, pp. 804–818, 2015.

[33] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, "Bayesian data analysis third edition (15 february 2021)," 15 February 2021.

[34] O. Martin, *Bayesian Analysis with Python: Introduction to statistical modeling and probabilistic programming using PyMC3 and ArviZ*. Packt Publishing Ltd, 2018.

[35] C. Davidson-Pilon, "Probabilistic programming and bayesian methods for hackers, 2013," *Unpublished. https://github. com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers*,

[36] C. Pilgrim, "Piecewise-regression (aka segmented regression) in python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3859, 2021.

[37] C. F. Jekel and G. Venter, "Pwlf: A python library for fitting 1d continuous piecewise linear functions," *URL: https://github. com/cjekel/piecewise_linear_fit_py*, 2019.

[38] V. M. Muggeo, "Estimating regression models with unknown break-points," *Statistics in medicine*, vol. 22, no. 19, pp. 3055–3071, 2003.

[39] V. M. Muggeo *et al.*, "Segmented: An r package to fit regression models with broken-line relationships," *R news*, vol. 8, no. 1, pp. 20–25, 2008.

[40] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *Proceedings of the 9th Python in Science Conference*, Austin, TX, vol. 57, 2010, pp. 10–25 080.

[41] M. DHAOUI, *Bayesian linear regression*, March 10, 2019. [Online]. Available: `https://mohameddhaoui.github.io/statistics/bayesianregression/#3--bayesian-linear-regression`.

[42] A. Gelman, "Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)," *Bayesian analysis*, vol. 1, no. 3, pp. 515–534, 2006.

[43] K. Burghardt, S. Guo, and K. Lerman, "Unequal impact and spatial aggregation distort covid-19 growth rates," *Philosophical Transactions of the Royal Society A*, vol. 380, no. 2214, p. 20 210 122, 2022.

[44] L. Zhang, J. Zhu, X. Wang, J. Yang, X. F. Liu, and X.-K. Xu, "Characterizing covid-19 transmission: Incubation period, reproduction rate, and multiple-generation spreading," *Frontiers in Physics*, vol. 8, p. 589 963, 2021.

[45] Y. Liu, R. M. Eggo, and A. J. Kucharski, "Secondary attack rate and superspreading events for sars-cov-2," *The Lancet*, vol. 395, no. 10227, e47, 2020.

[46] M. DHAOUI, *Research article | open access. volume 2021 |article id 6927985*. [Online]. Available: `https://doi.org/10.1155/2021/6927985`.

[47] M. B. Perry, *The exponentially weighted moving average*. [Online]. Available: `https://www.researchgate.net/publication/313992620_The_Exponentially_Weighted_Moving_Average`.

[48] X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012.

[49] S. Prabhakaran, *Augmented dickey fuller test (adf test) – must read guide*. [Online]. Available: `https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/`.

[50] M. H. Rahman, U. Salma, M. M. Hossain, and M. T. F. Khan, "Revenue forecasting using holt–winters exponential smoothing," *Research & Reviews: Journal of Statistics*, vol. 5, no. 3, pp. 19–25, 2016.

[51] S. Srihari, *Long-short term memory and other gated rnns*. [Online]. Available: `https://cedar.buffalo.edu/~srihari/CSE676/10.10%5C%20LSTM.pdf`.