Research paper on Forecasting the spread of covid-19 in Canada using Machine Learning models

Machine Learning (BITS F464) - Ashwin Srinivasan,Tirtharaj Dash
Department of Computer Science & Information Systems,

BITS Pilani, K.K. Birla Goa Campus

30 April 2021

Team members :

Rakshit - 2019A3PS0510G

Vinay R Sabarad - 2019A4PS0788G

Arka Nayak - 2018A7PS0159G

Shanmukha Aditya V - 2018A7PS0688G

Table of Contents

-

ABSTRACT

The COVID-19 pandemic is an ongoing global pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The virus was first identified in December 2019 in Wuhan, China. The World Health Organization declared a Public Health Emergency of International Concern regarding COVID-19 on 30 January 2020, and later declared a pandemic on 11 March 2020. As of 30 April 2021, more than 150 million cases have been confirmed, with more than 3.16 million deaths attributed to COVID-19, making it one of the deadliest pandemics in history.

In the absence of any specific effective treatment for COVID-19, Governments and health organisations around the world resorted to strict preventive measures ranging from wearing masks in public places and social distancing to complete country wide lockdowns and closing of international borders. Also swift vaccine development and administration gave a ray of hope in the fight against the novel coronavirus. But the second wave of the pandemic has brought back panic once more and administrations around the world are struggling to cope up with it.

This paper presents our study of the COVID-19 pandemic in Canada. Visualizations of various measurable attributes related to the pandemic gave us considerable insights into their interdependencies and the effects of government measures on the spread of the infections. We used Machine Learning models to learn about the impact of lockdown and vaccination on the case numbers and used it to predict an effective measure to control the second wave of COVID-19 in Canada.

INTRODUCTION

Machine learning provides a lot of support in identifying the disease with the help of image and textual data. Machine learning can be used for the identification of novel coronavirus. It can also forecast the nature of the virus across the globe. However, machine learning requires a huge amount of data for classifying or predicting diseases. In this process, main data which plays a crucial role is number of cases, number of recoveries, number of deaths, and number of vaccinated.

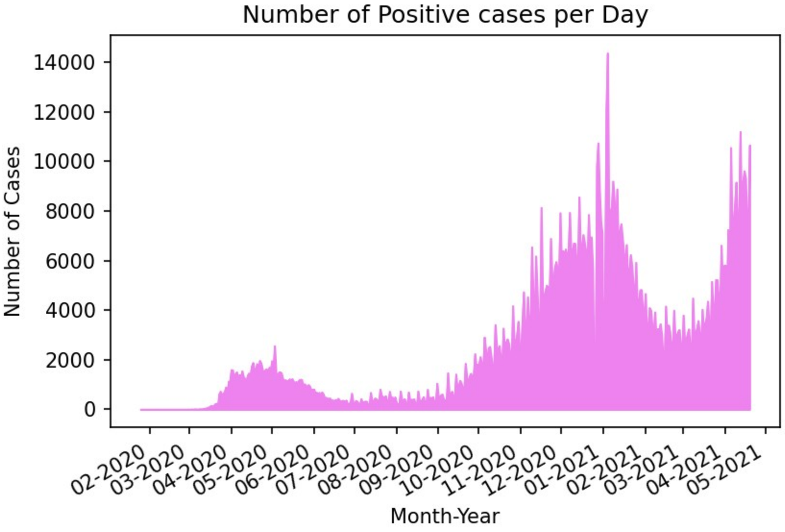
We have used Python libraries (Matplotlib and Seaborn) to visualise the available data and Machine Learning models (Support Vector Regression, Linear Regression and Polynomial Regression) to forecast the spread of Covid-19 and a mathematical model (Susceptible Infected and Recovered) to analyse spread of Covid-19.

This paper studied how machine learning algorithms and methods can be employed to fight the COVID-19 virus and the pandemic. It further discusses the primary machine learning methods that are helpful during the COVID-19 pandemic. We further identified and discussed algorithms used in machine learning and their significant applications. The study is aimed to investigate and assess the effectiveness of preventive measures to be taken by government of Canada to control the spread of Covid 19. It also explains mathematical prediction of covid cases of Canada for a short period.

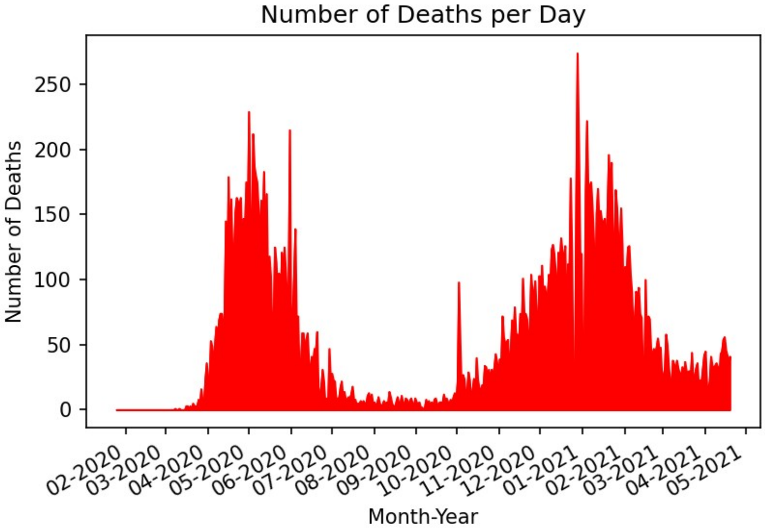
DATA VISUALIZATION AND PRIMARY INFERENCES

By using visual elements like charts, graphs, and maps, data visualizationtools provide an accessible way to see and understand trends, outliers, and patterns in data.Visualizing various parameters related to the pandemic helped us in identifying key trends and also the causes and effects of government measures.

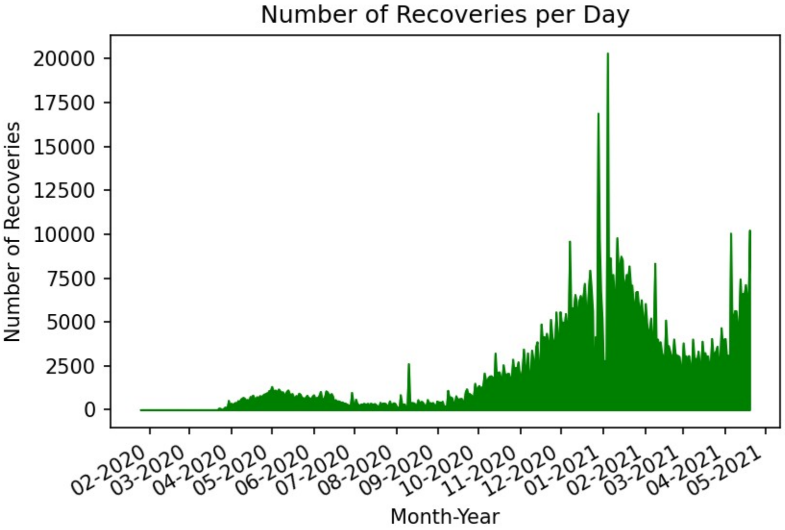
3.1) We can see three peaks in the cases curve. Thus unlike other regions Canada can be thought of as having a third wave.



3.2)We can see two valleys in the death curve. The first valley can be attributed to the lockdown related restrictions and the second one due to the vaccination drive.



3.3)The number of recoveries have been increasing which is a positive sign.



3.4)Moving averages is a useful technique to analyse the trend in a time series data. Here we have used the weighted moving average i.e. the most recent data point is give more weightage and the weights decrease backwards. We have used the moving average crossover between the 25-DAY WMA AND 50-DAY WMA to analyse the important events that might influence the trends.From the figure we can conclude that:

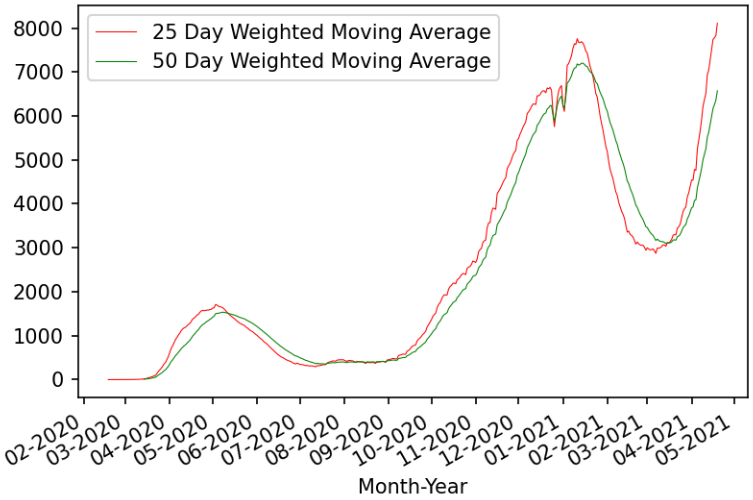
3.4.1) Initially from March to May 2020 there was an continuous and widespread increase in Covid cases as the short term moving average is higher than the long term average.

3.4.2) Around May-June 2020 the short term average crosses below the long term average and cases are decreasing. This is exactly when strict Covid restrictions were announced in Canada. Thus we can say Government restrictions led to a decrease in cases.

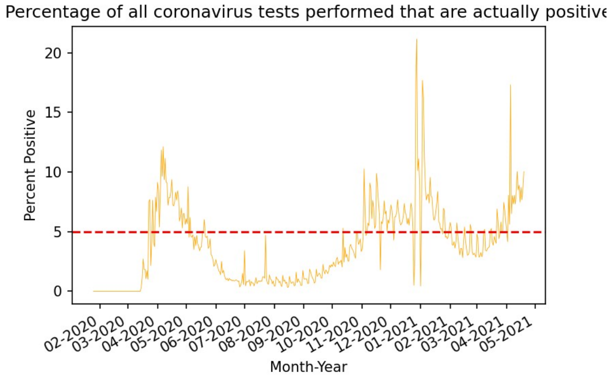
3.4.3) Around August-September 2020 the short term average crosses above the long term average with an increase in cases which can be mainly attributed to increase in cases after covid restrictions were eased in Canada.

3.4.4) Around January 2021 the short term average again crosses below the long term average. This can be mainly attributed to some provicewise restrictions and also widespread vaccination.

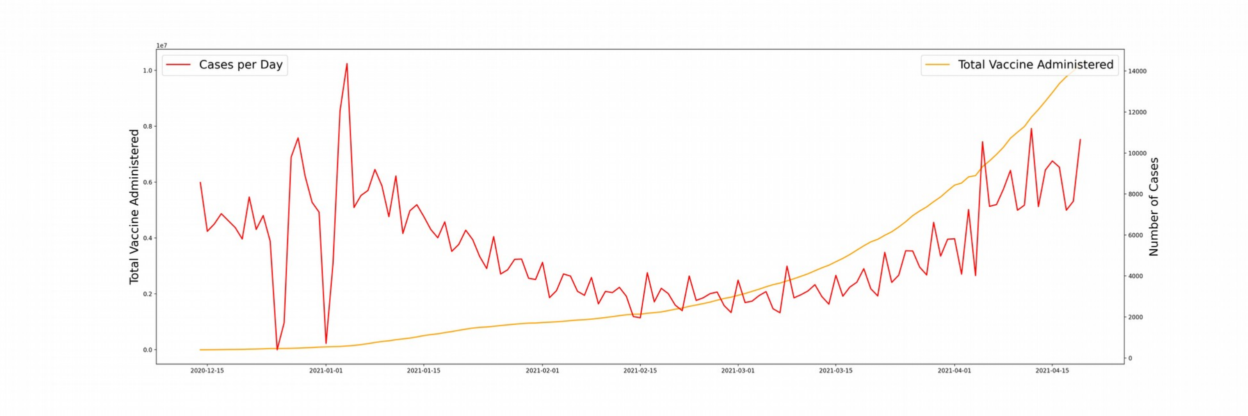
3.4.5) Around March-April 2021 the short term average crosses above the long term average indicating a new wave of Coronavirus infections.



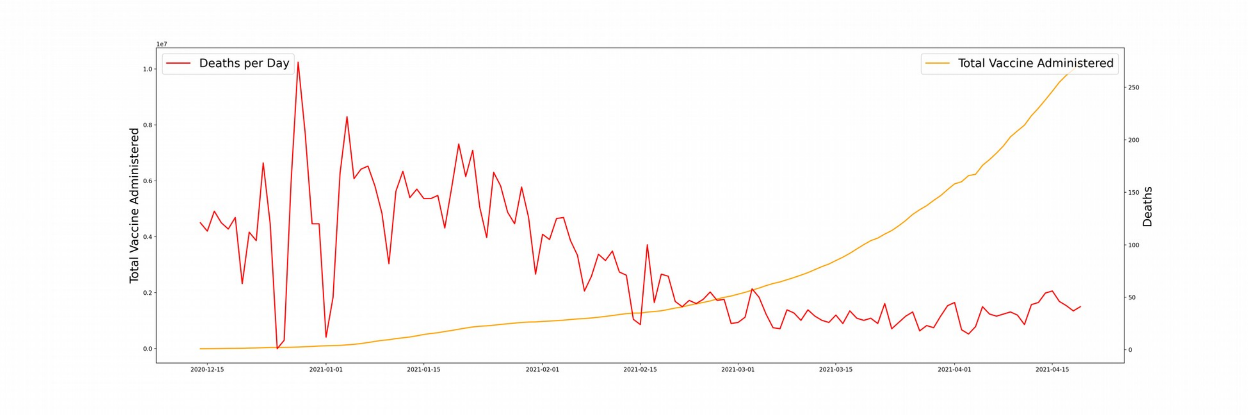
3.5) The percent positive is the percentage of all coronavirus tests performed that are actually positive, or: (positive tests)/(total tests) x 100%. The percent positive (sometimes called the “percent positive rate” or “positivity rate”) helps public health officials answer questions such as: What is the current level of SARS-CoV-2 (coronavirus) transmission in the community? Are we doing enough testing for the amount of people who are getting infected? The percent positive will be high if the number of positive tests is too high, or if the number of total tests is too low. A higher percent positive suggests higher transmission and that there are likely more people with coronavirus in the community who haven’t been tested yet. the World Health Organization recommended in May that the percent positive remain below 5% for at least two weeks before governments consider reopening. Canada followed these rules but the cases began rising again after restrictions were eased and the positive percent have been >5% since then.


Percentage of all coronavirus tests performed that are actually positive

3.6) As can be seen with an increase in vaccination there was a decrease in the number of cases initially but the case count started increasing thereafter. Thus we cannot conclude whether vaccination can lead to decrease in the number of infections spreading .



3.7) It is clear from the graph that the increase in vaccination and certainly led to a decrease in the deaths. Thus widespread vaccination can be concluded to be an effective measure to curb the Covid related deaths.



MODELS IMPLEMENTED IN THE CASE STUDY

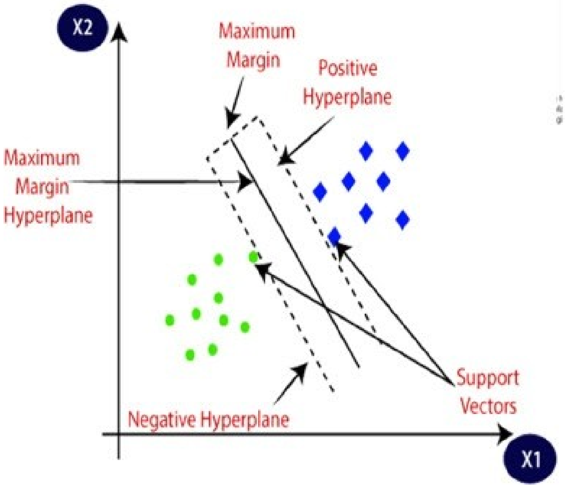
The data collected in the case study was analysed using various machine learning models to predict the rise in number of cases, deaths and vaccination. The impact of vaccination was quantified.

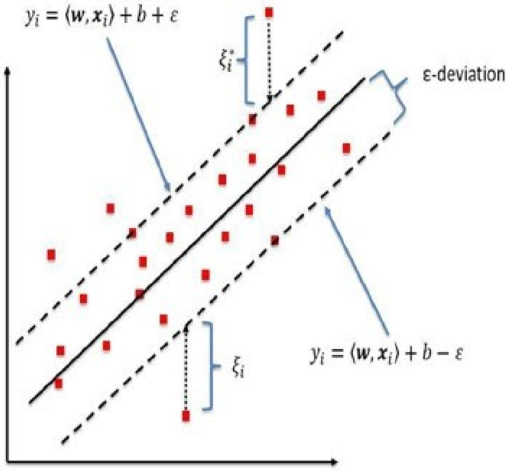
4.1) Machine Learning Models Implemented in the case study :

4.1.1)SVR (SVM Regression) :

This study uses the Machine Learning models to forecast the number of upcoming cases affected by COVID-19. Particularly two models for forecasting such as Support Vector Regression (SVR) and Linear regression (LR) are used in this study to predict the future number of cases of COVID-19. Number of upcoming covid-19 cases are going to happen and deaths are predicted by using these two models for the next 70 days. The predictions are done for Canada. Between these two SVR is performing well compared with Linear Regression.

The SVM algorithm is versatile, it supports linear , nonlinear classification and regression. We have used the SVR to bound the data between an upper and lower limit and predict the number of active cases , recovered cases and deaths in the next 20 days.We can think of Support Vector Regression as the counterpart of SVM for regression problems. SVR acknowledges the presence of non-linearity in the data and provides a proficient prediction model.



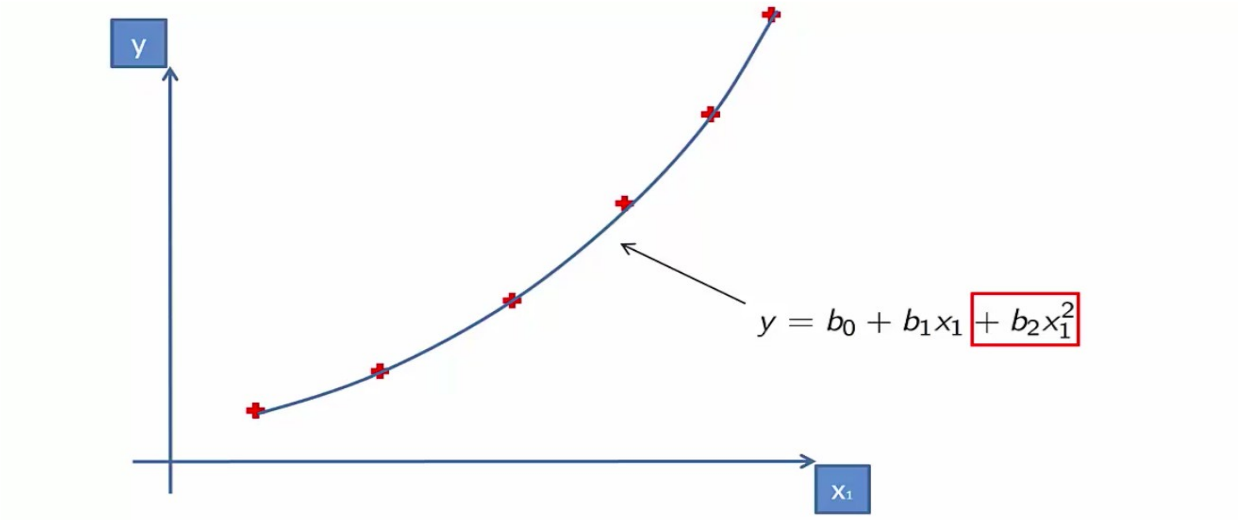

$y_i = \langle w, x_i \rangle + b + \varepsilon$
ξ_i^*
ε-deviation
ξ_i
$y_i = \langle w, x_i \rangle + b - \varepsilon$

4.1.2)Polynomial Regression :

A polynomial term turns a linear regression model into a curve but it still qualifies as a linear model. The polynomial models quadratic, third-degree, fourth-degree, fifth-degree, and sixth-degree were used in those situations. The nth order polynomial model in one variable is given by the equation :

$$y = a_1 . X + a_2 . X^2 + a_3 . X^3 + a_3 . X^4 + ... + a_n . X^n + \varepsilon$$

where (n= 2, ….,6) represents the degree of the models.

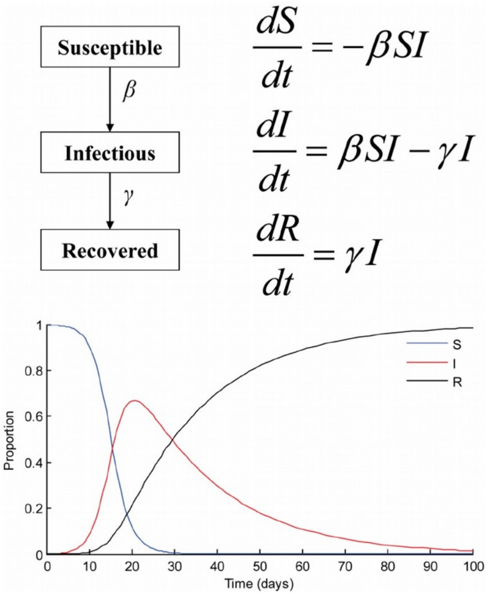


Disease diffusion is defined as the cumulatively increasing degree of spread of a particular disease among humans or animals from a region of outbreak to other regions, until the disease has spread across all regions. So, models developed for studying transmission processes of infectious diseases theoretically can be termed as diffusion models. Thus, future development trends of infectious diseases can be accurately predicted. Thus, in this research, we fitted hierarchical polynomial regression models on daily cases of COVID-19 globally and forecast new cases , deaths and recovered cases in the upcoming days.

4.2)Mathematical Model:

4.2.1)SIR (Susceptible Infected and Recovered):

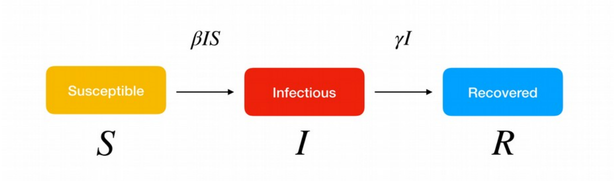
As the first step in the modeling process, we identify the independent and dependent variables. The independent variable is time **t**, measured in days. We consider two related sets of dependent variables.The first set of dependent variables counts *people* in each of the groups, each as a function of time:



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

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

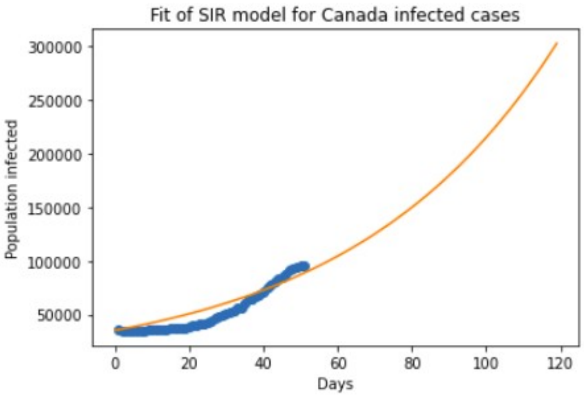
$$\frac{dR}{dt} = \gamma I$$

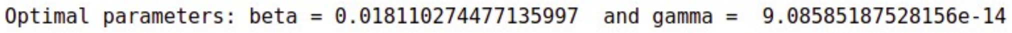


SIR function:

Moving from the susceptible to the infected cases, the contact rate, β, determines the disease velocity in the population. In detail, the transition from S-state to I-state is not deterministic, but is always stochastic. Hence, β includes multiplication of rate, probability, and population number.

Using this model we have predicted the number of infected persons for the next few days based on pevious fifty days.We optimised and got values of beta and gamma as 0.018111 and 9.08585e-14 respectively.





```
Optimal parameters: beta = 0.018110274477135997  and gamma =  9.08585187528156e-14
```

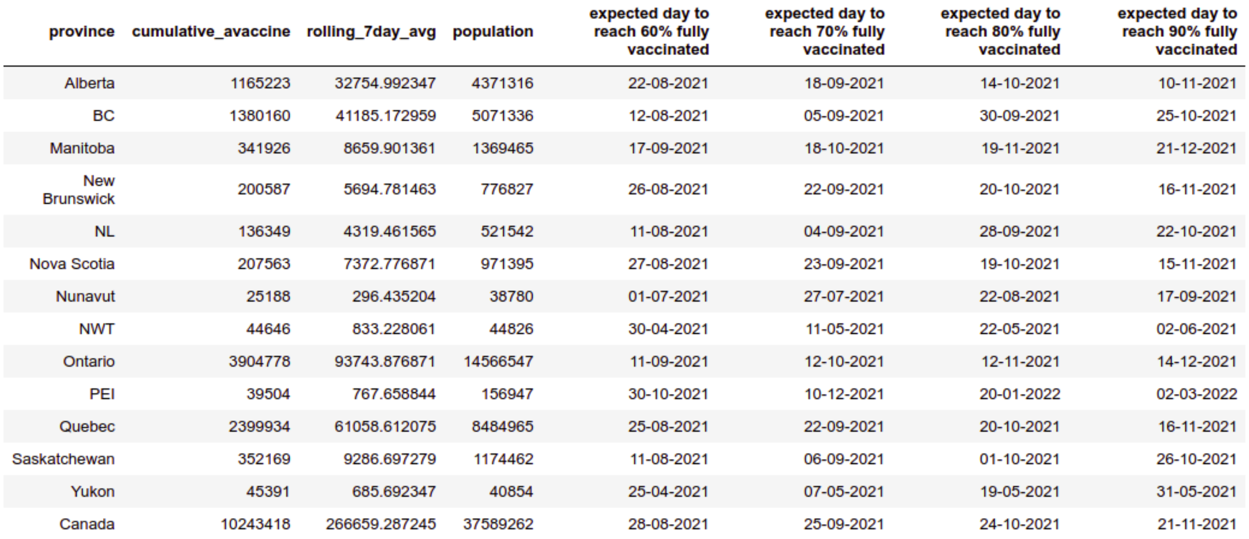
As it is theoritical model,assuming no other factors involved the number of infected persons are predicted as above.But in all our models polynomial regression showed less error than others.

VACCINATION, SOCIAL DISTANCING AND RESULTS

5.1)Vaccination:

From the data collected we have forecasted the active cases,mortality cases,recovered cases.Wehave also predicted the approximate date of fully vaccinated percentage using vaccine_administration data. We have calculated the moving average of 7 days for vaccine_administration and took the average for the last 2 weeks.Following is the predicted dates for fully vaccinated percentage and their corresponding rolling averages,population.

We calculated the same for provinces in Canada.From our prediction,We expect Canada to be 70% fully vaccinated by 25th sep,2021 and 80% fully vaccinated by 24th oct 2021.We expect Prince Edward Island(PEI) to be fully vaccinated much late due to low rolling 7-day average for vaccine administration when compared to other provinces.

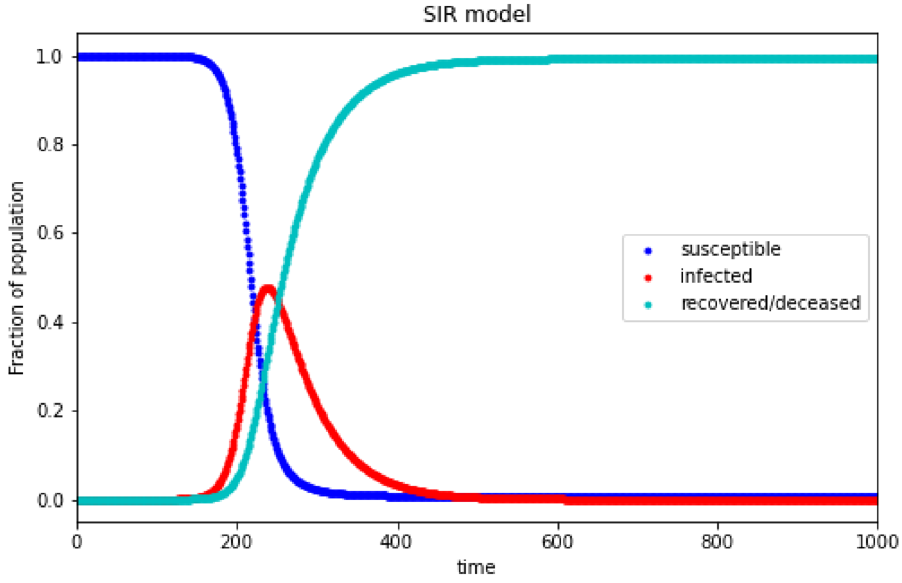


province	cumulative_avaccine	rolling_7day_avg	population	expected day to reach 60% fully vaccinated	expected day to reach 70% fully vaccinated	expected day to reach 80% fully vaccinated	expected day to reach 90% fully vaccinated
Alberta	1165223	32754.992347	4371316	22-08-2021	18-09-2021	14-10-2021	10-11-2021
BC	1380160	41185.172959	5071336	12-08-2021	05-09-2021	30-09-2021	25-10-2021
Manitoba	341926	8659.901361	1369465	17-09-2021	18-10-2021	19-11-2021	21-12-2021
New Brunswick	200587	5694.781463	776827	26-08-2021	22-09-2021	20-10-2021	16-11-2021
NL	136349	4319.461565	521542	11-08-2021	04-09-2021	28-09-2021	22-10-2021
Nova Scotia	207563	7372.776871	971395	27-08-2021	23-09-2021	19-10-2021	15-11-2021
Nunavut	25188	296.435204	38780	01-07-2021	27-07-2021	22-08-2021	17-09-2021
NWT	44646	833.228061	44826	30-04-2021	11-05-2021	22-05-2021	02-06-2021
Ontario	3904778	93743.876871	14566547	11-09-2021	12-10-2021	12-11-2021	14-12-2021
PEI	39504	767.658844	156947	30-10-2021	10-12-2021	20-01-2022	02-03-2022
Quebec	2399934	61058.612075	8484965	25-08-2021	22-09-2021	20-10-2021	16-11-2021
Saskatchewan	352169	9286.697279	1174462	11-08-2021	06-09-2021	01-10-2021	26-10-2021
Yukon	45391	685.692347	40854	25-04-2021	07-05-2021	19-05-2021	31-05-2021
Canada	10243418	266659.287245	37589262	28-08-2021	25-09-2021	24-10-2021	21-11-2021

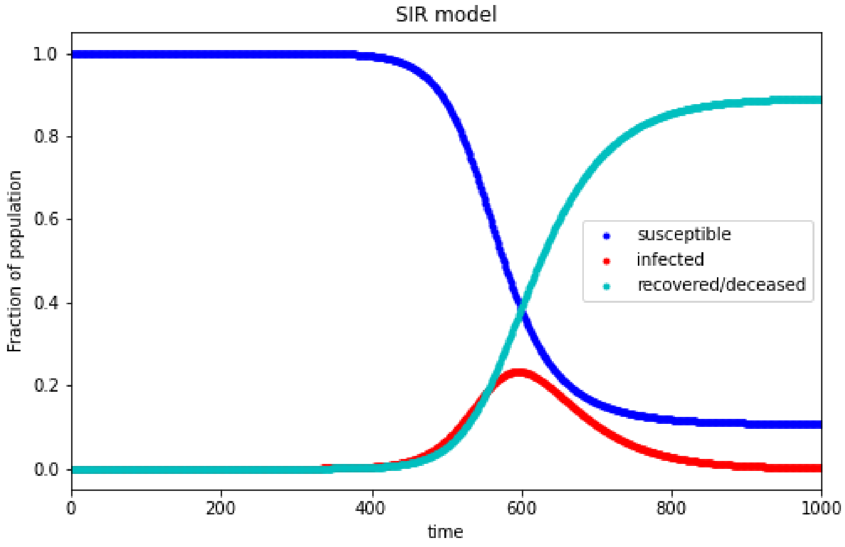
5.2)Social Distancing:

We have analyzed the social distancing role in covid infection rate.We used the mathematical model SIR to show its role by plotting infected cases and how social distancing can flatten the curve by decreasing “beta”.

5.2.1)In the first plot if we avoid social distancing, we take beta as 1 and the fraction of the infected population is higher.



5.2.2) Whereas in the next one, if we maintain social distancing,the infection rate will be less when compared to above one.So we take beta as 0.5 and the fraction of the infected population is much lower when compared to the one above.



CONCLUSION

A forecast of COVID-19 spread in Canada was carried out using various statistics and machine learning modeling approaches. The forecast is based on the data from 15 February 2020 until 19 April 2021. It's aim is to investigate and assess the effectiveness of preventive measures of the government of Canada to control the spread of COVID-19. According to our analysis, the data of the Covid-19 cases is non linear and imposing the lockdown has an impact on the rate of growth in the number of cases. On imposing strict restrictions, it is observed that the rate of growth of Covid-19 cases decreases whereas when the restrictions are eased, then it is observed that the rate of growth of Covid-19 cases increases. It is observed that although we can not predict the relation between vaccination and number of daily cases, but it can be clearly seen that vaccination has decreased the number of deaths.

These models also predicted the outbreak of the COVID-19 in Canada for the next 20 days, 100 days, the final size of the infected cases, and the final time of the epidemic. We have found out that the best of the proposed models namely exponential sixth-degree polynomials are strong residual and prediction for the next 20 days.These models are very useful for the Canadian government for managing the COVID-19 outbreak for the next few months.

WORKS CITED

- https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_Canada
- https://www.worldometers.info/coronavirus/country/canada/
- https://github.com/ccodwg/Covid19Canada
- https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Canada
- https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7321055/