Worldwide differences of Covid-19 on cases and deaths using time series forecasting models

MSc Research Project
Data Analytics

Ankish Kumar Chandani
Student ID: x18189245

School of Computing
National College of Ireland

Supervisor: Prof. Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ankish Kumar Chandani
Student ID:	x18189245
Programme:	MSc. in Data Analytics
Year:	2019-20
Module:	MSc Research Project
Supervisor:	Prof. Christian Horn
Submission Due Date:	28/9/2020
Project Title:	Worldwide differences of Covid-19 on cases and deaths using time series forecasting models
Word Count:	7296
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	27th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Worldwide differences of Covid-19 on cases and deaths using time series forecasting models

Ankish Kumar Chandani
x18189245

Abstract

Millions of people have been infected and hundreds of thousands have died due to a deadly disease Covid-19. Hence, it has become a significant research area since it started spreading around the world. Covid-19 forecasting demands accurate reported count to analyze time-related data along with complex model implementation. Accurate Covid-19 forecasting is critical for planned action to combat this disease and take necessary precautionary measures to reduce its impact. The United States has been ranked as the most impacted country in the world and Italy have the lowest mortality rate among the countries infected with more than 150,000. Implementation of advanced forecast models can help in acquiring valuable future forecasting which could aid governments as well as health organizations to work together and guide the public to help prevent this virus. A novel forecast model, Prophet[1] has been implemented in this research to predict and forecast future Covid-19 cases and deaths precisely. Comparison of model error and forecast data was performed for the machine learning methods i.e. Polynomial Regression, Holt's Linear Model, and time series such as AR, ARIMA model. Evaluation metrics such as MAPE, RMSE and MAE have been used to evaluate the model performance. Research finding signifies the most efficient model to be ARIMA and Prophet. ARIMA and Prophet model gave a better performance in predicting and forecasting the total Covid-19 cases and deaths respectively along with the lowest combined evaluation errors. This approach can assist the governments to put necessary regulations in place before millions more get infected and also prevent the loss of billions worth of money.

Index Terms: Covid-19, Prediction, Forecasting, Time Series, ARIMA, Prophet

1 Introduction

A bunch of patients suffering from pneumonia of unknown etiology were admitted in hospitals of Wuhan, China in late December 2019 and it became an outbreak within weeks when the total number of cases and fatalities surpassed Severe Acute Respiratory Syndrome (SARS) cases and deaths. Initially, this disease was named as a 2019-novel coronavirus (2019-nCOV) as it is caused by a novel beta coronavirus but later, this disease was officially named as coronavirus disease 2019 (Covid-19) by World Health Organization (WHO) and the new coronavirus was officially named as SARS-CoV2. The human

[1] `https://facebook.github.io/prophet/`

respiratory system is primarily targeted by this coronavirus and became a global health threat. Even the previous coronaviruses outbreaks of SARS and the Middle East Respiratory Syndrome (MERS) were the global health threat but as of date, it is the largest outbreak related to typical pneumonia. The admitted patients of Wuhan were found to be connected with the exposure to seafood and linked to wet animal market (Nishiura et al.; 2020). As of 30th Jan 2020, the number of cases continued to exponentially escalate and spread across 34 regions of China and the same day, WHO declared Covid-19 to be a concern of international public health emergency (Mahase; 2020). Below figure 1 shows the precautions and symptoms of Covid-19[2]

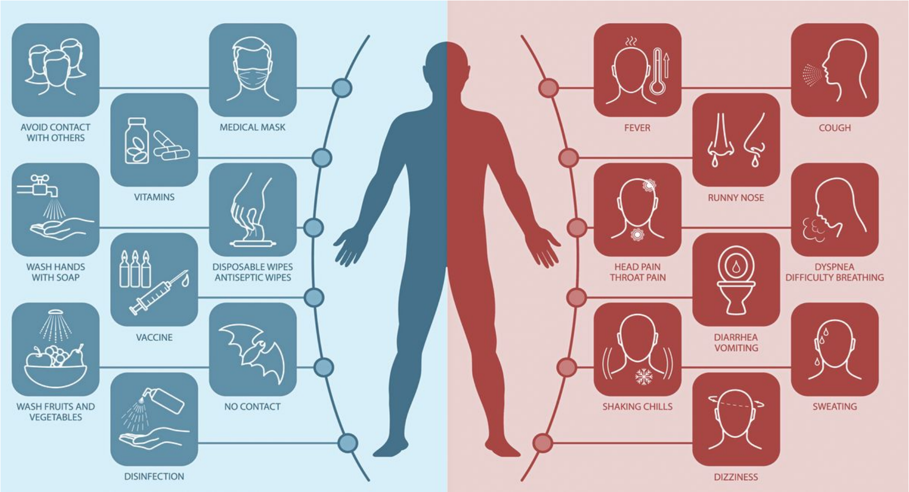


Figure 1: Covid-19 Precautions and Symptoms

Despite the lockdown measures in place and the suspension of flights, trains, and even public transport worldwide, the confirmed cases were around 20 million and total deaths were more than 700k by 10th August 2020 (Organization et al.; 2020). However, there has been still a debate on the transmission of Covid-19 from intermediate hosts like bats to humans. Although it has been confirmed that there is a human to human transmission by multiple means namely aerosols, droplets, and fomites (Wang and Du; 2020), as an increasing number of patients have not been reported to be exposed to the wet animal market and cases have been reported for healthcare workers. The person with Covid-19 could be triggered with mild symptoms of pneumonia. Fever, cough, headache, sore throat, coryza, fatigue, breathing difficulty, vomiting, myalgia (Chen et al.; 2020) are the symptoms of Covid-19. Viral particles can be spread by an infected person while talking, breathing, coughing, or sneezing. Older people and people with pre-medical conditions are more vulnerable to catch the virus as compared to others. The maximum number of cases and fatalities have been reported in the United States of America. However, there has been negligence in using the face masks which is highly important to stop the transmission. Social distancing has been promising till date in restricting the virus and most of the governments have implemented a rule of 2m gap distance between people in the public places.

Due to the shortage of test kits and hospital beds, there has been pressure on the governments to tackle this problem as quickly as possible to save more lives. More than 200 countries have been affected by this deadly virus and it is important to analyze its

[2] https://www.vox.com/2020/4/2/21200217/coronavirus-symptoms-covid-19-fever-cough-smell-taste

impact to prevent the virus from spreading. As of 11th August 2020, there has been no vaccination available but multiple pharma companies are trying hard in their capacity to develop the vaccination which could help prevent this deadly virus to spread. Several prediction models have been implemented using the combination of different parameters or variables to estimate the risk of getting infected.
This paper focuses on the prediction and forecasting of the total cases and deaths due to Covid-19 worldwide. For this research, time-series data of Covid-19 will be analyzed. Different time series Forecasting models will be used and the performance of these models will be evaluated using different metrics.

1.1 Motivation

Millions of people have been affected and hundreds of thousands of lives have been lost due to Covid-19. People have lost their loved ones, so many people lost their jobs and till date, people are still struggling to come out of that fear of getting infected by shaking hands, going out together, meeting friends outdoor or having food in the restaurant. If infected, mild illness can be the reason of death. Nobody is safe in the whole world now, even top ministers and wealthy people are getting infected. This research can help the government and common people to learn about the impact of various factors on Covid-19 and forecast the total cases and deaths which might help the government to take necessary precautions which will lead to saving plenty of lives.

Research Objective: Analysis and future impact of Covid-19 on total confirmed cases and total deaths worldwide.

The rest of the paper is organized as follows: Section 2 presents the related work. Section 3 shows the research methodology, analysis and pre-processing. Section 4 talks about the models implemented and evaluation on the time series data. Discussion about the evaluation is presented in section 5, Conclusion and Future work of the research is discussed in section 6 and acknowledgement in the last section.

2 Related Work

2.1 Statistical Analysis

A systematic study was conducted as a meta-analysis, which included 1576 infected cases who were laboratory-confirmed in China (Yang et al.; 2020). Risk of various comorbidities in critical patients compare to non-critical patients was calculated by the odds ratios (ORs) along with the estimate of the 95% confidence intervals (CIs). Fever was found to be the most prevalent medical symptom followed by cough, fatigue and dyspnea. Hypertension with 21.1% prevalence was the most prevalent one in critical cases followed by diabetes, cardiovascular disease and respiratory disease. Age and comorbidities were found to be the major risk factor for critical patients (the majority of them were older). However, this study did not find high heterogeneity statistics as well as subgroup analysis and sensitivity analysis. Some patients were still in the hospital. Flu vaccine could help decrease influenza and pneumonia incidence with diabetes people aged ¡65 years by 43%, and people aged 65 years by 55%. The symptoms of influenza (cough, fever, or fatigue), and Covid-19 were quite similar.

41 patients admitted in the hospital were laboratory confirmed as Covid-19 infections in Wuhan, China by 2nd Jan 2020. Out of these 41 patients, 20 belonged to the age group 25-49 years and 14 were in the age group 50-64 years (Huang et al.; 2020). The median age of the infected patient was found to be 49 yrs. There was no infection in the children and 30 (73%) patients were males, out of which some already had underlying diseases including hypertension, cardiovascular disease, and diabetes. The most common symptom was fever, cough, and fatigue. All patients were suffering from pneumonia with abnormalities in chest CT. The average time from symptoms to hospital admission was 7 (4-8) days, to shortness of breath was 8 (5-13) days, to ARDS was 9 (8-14) days, and 10.5 days to ventilation and ICU admission. 27 of 41 patients were exposed to Huanan seafood market which indicated Covid-19 started from this market. However, the author was only able to experiment on 41 patients, it was not possible to evaluate the host health risks with multivariable-adjusted methods for disease severity and mortality. Also, there was no adolescent or paediatric patients may be due to potential risk bias in this cohort study.

(Zhong et al.; 2020) used a mathematical model, Susceptible-Infectious-Recovered (SIR) model, a dynamical epidemic model to present the early prediction of the Covid-19 epidemic of mainland China. Epidemiological data were first analyzed to obtain an approximate estimation of the infection rate and then based on infection and removal rate, multiple experiments were performed to predict Covid-19 spread using medical care and anti-epidemic measure. The population was divided into three classes Susceptible (S), Infectious (I) and Recovered (R) and infection rate and removal rate were obtained statistically. The model showed that after day 10, Covid-19 showed a pulse-like increase in the infected cases. Between 11-15 Jan, there were 41 cases whereas the infected cases were almost doubled on 18th Jan, which was the reason for pulse-like infection increase of 1.44 / day. This sharp increase might be the result of diagnosis technique and China started updating the epidemiological data. The model predicted the number of unrecovered infections to reach a peak value of 43,000 cases from late February to early March. Also, the number of infected cases expected to rapidly decrease when the epidemic fade out in June. Overall, the author predicted Covid-19 to persist from three to five months and final cumulative infected cases to be around 140,000. If the government provide good medical support and enhance hospitalization capacity, then cumulative cases could go down to 80,000 and the severity of Covid-19 may be at low-level.

Wang, Tang and Wei (2020) conducted statistical analysis for the period 1st Dec 2019 to 9:30 am 26th Jan 2020 of mainland China and found mortality to be approximately 2.84% with 1975 confirmed cases and 56 deaths. The median death age was found to be 75 years. Fever and cough were the most prevalent early stages of deaths with 64.7% and 52.9% respectively. The average duration from the first diagnosis to death was 14 days ranging between 6-41 days and seemed to be less for people i.e. 11.5 days who are aged 70 and above compare to people aged below 70 years i.e. 20 days. The infection quickly increased due to Spring Festival travel rush between 20th Jan 2020 and 25th Jan 2020 and the mortality of Covid-19 (2.84%) during that period was lower than SARS-CoV (9.6%) which had transmitted globally between November 2002 to July 2003 and killed 774 people and lower than MERS-CoV (34.4%) mortality which had also spread to 27 countries globally Sept 2012 to Sept 2019 killing 858 persons. Even though the mortality of Covid-19 was lower comparatively, but it seemed to be highly contagious. The study also concluded that the elderly people died first and most of them were already suffering from one of the other comorbidities before admission however the relationship between

deaths due to Covid-19 and underlying medical conditions were still not clear.
(Fanelli and Piazza; 2020) analyzed the outbreak in 2 countries, Italy and China between 22nd Jan to 15th Mar using the SIR model and concluded that irrespective of these three countries, the recovery rate i.e the rate at which people are being recovered was almost same, but the rate of confirmed cases and deaths were more variable. Around 21st March, Italy would be the peak of infected cases around 26000 excluding recovered and death cases while till the end of this pandemic, death cases could reach 18k. Morality rate could be between 4-8% where in China, it would be 1-3%. There would be a need of approximately 2.5k ventilators by health authorities in Italy to tackle the epidemic. Without the implementation of lockdown measures, the Covid-19 could be at the peak between Feb end and Mar start in Wuhan, China and the second wave was likely to happen if the lockdown measures and restrictive control is lifted (Roda et al.; 2020). Another research (Wang, Li, Guo, Xie, Yao, Cao, Day, Howard, Graff, Gu, Ji, Gu and Sun; 2020) predicted and forecasted the mortality rate using the Patient Information Based Algorithm (PIBA) to determine whether there is a need of public attention or not. It takes approximately 13 days for showing symptoms of death in China. However, the prediction might vary with the actual due to skewed and limited data. Many people who were suffering from mild illness were not admitted and is the reason for the mortality rate to be high.
Italy had 12462 confirmed cases and 827 deaths as of 11th March 2020 and almost two-thirds of these patients were already suffering from diabetes, cancer, or cardiovascular diseases, or were former smokers (Remuzzi and Remuzzi; 2020). However, the patients who were diagnosed with acute respiratory distress syndrome (ARDS) caused by respiratory syndrome coronavirus 2 pneumonia needed respiratory support otherwise they would die. Between 1-11 Mar 2020, the number of intensive care patients registered daily in Italy was regularly between 9% and 11% of patients who were actively contaminated. An aggressive and effective approach was required to take care of critically ill patients often requiring ventilator support system. Wuhan city and other province had heterogeneity in the transmission between each other. The rising pattern required more than 2500 hospital beds in a week time to handle Covid-19 patients. The most successful way of suppressing this virus epidemic was possibly to prevent near interaction at the person level and social gatherings in-region. The accuracy between recorded evidence and exponential estimation was very similar until day 17 and more 30k patients could be infected by 15 March if the same pattern continued.
There was an urgent need for developing methods to identify different transmission modes such as urine and faecal samples (Rothan and Byrareddy; 2020) to formulate plans for minimizing as well as inhibiting transmission and developing disease prevention therapies. There was an uncertainty about pregnant women transmitting the disease to the child as they were extremely vulnerable to pneumonia and respiratory infections. As per the phylogenetic reports and genetic sequence identity, Covid-19 was extremely distinctive from SARS-CoV and thus may be deemed a modern beta coronavirus infecting human. The spread of Covid-19 infections among individuals led to patient isolation despite a range of therapies. There were no potential antiviral antibiotics or vaccination against the infection for future human therapy. Decontaminating hand wash reagents was required in public utilities and hospitals regularly. Nonetheless, more research was required urgently to classify novel chemotherapeutic therapies of coronavirus infections. There have been multiple epidemic models for prediction and a study (Farahi and Kamandi; 2020) proposed a dynamic epidemic reproductive model. Constant reproductive number (R0) was not enough to predict the correct number of infected people, so a dynamic reproductive

number(R0) was required for predicting the spreading of the virus. Dynamic R0 helped to compare the predicted infected cases versus the real infected cases till 31st Jan 2020 and the transmission rate. Performance of the prediction model was high and kept the public aware of the virus so that maximum precautions could bring a big change in the result of the epidemic.

2.2 Machine Learning Algorithms

As of 11th May 2020, Total Covid-19 cases and total deaths in India were 67,161 and 2,212, respectively. Few testing kits were available in the hospitals which were not sufficient for testing the rapidly increasing cases. No defined treatment was in place for Covid-19 infected patients.(Kolla; 2020) proposed various machine learning algorithms to predict the age group highly vulnerable to Covid-19. Random Forest Classifier model outperformed and gave better accuracy and coefficient of Determination than other models such as SVM, Logistic Regression, Decision tree Classifier. Most of the infected patients were of the age group 20-50 years. (Arun and Iyer; 2020) proposed multiple learning algorithms such as Bayesian Ridge, Support Vector Machine (SVM), SIR model, RNN and Po1ynomial Regression to predict the pandemic scale, fatality rate and recovery rate as well as Covid-19 transmission. RNN outperformed other machine learning algorithms with better accuracy and precision whereas outliers were affected in Polynomial Regression and data overfitting was caused in SVM. The analysis concluded the pandemic to be deadly and highly infectious as compare to patient's recovery rate.

A hybrid artificial-intelligence (AI) model to predict Covid-19 trend was proposed by first using an improved susceptible-infected (ISI) model and then embedding natural language processing (NLP) along with long short-term memory (LSTM) for several typical cities and provinces in China (Zheng et al.; 2020). Cases increased till 23rd Jan 2020 but as soon as lockdown measures were put into action from 24th Jan, it played a critical role in the cases declining in Wuhan (Covid-19 epicentre). The model predicted the cumulative confirmed cases to be around 48k by the end of March but if the lockdown measure had been implemented on 27th Jan instead of 24th then the cases in Wuhan would have been close to 102k. The proposed model ISI+NLP+LSTM achieved better precision prediction than ISI or ISI+LSTM model and obtained mean absolute percentage errors (MAPE) with 0.05%, 0.38%, 0.52% and 0.86% for Shanghai, Beijing, Wuhan and worldwide respectively. Transmission of human-to-human has to be strictly limited to reduce infections especially among healthcare workers and further implementing the travel restrictions on international people coming from China to help control the spread (Lai et al.; 2020). This transmission is spread via direct contact or droplets and the infection period of Covid-19 is found to have an estimated incubation mean of 6.4 days and a reproduction number between 2.24-3.58. Fever followed by cough was the most common symptom among patients diagnosed with pneumonia caused by a coronavirus.

In the absence of Covid-19 treatment and vaccination, it was necessary not to only focus on the complete elimination but qualitative disease control.(Mandal et al.; 2020) analyzed and predicted the cumulative infected person in three states of India (Maharashtra, Delhi, and Tamil Nadu) till 27th April 2020 using mathematical analysis. India being the 2nd most populated country in the world is the 3rd most affected country by Covid-19 as of 20th July 2020. Government policing was important in controlling the nature of this destructive disease. Short term analysis is only possible and can be accurate due to the frequent changes in government policy. Determining the reproduction number could help in the disease control measure. Delhi and Maharashtra were found to have an increasing trend even with proper government measures in place while Tamil Nadu was estimated to be in control with existing space of parameter. People not taking this disease seriously, high density of population and people coming out of their home frequently even for essential material has been the major burdens behind Covid-19 control. Proper implementation of public hygiene, lockdown measures and people following quarantine could play an important role in combating the transmission of Covid-19. There was a need for the local administration to conduct a screening for the patients quarantined from the past 14 days. Four states in India were the most affected and required high attention from authorities.(Bhati and Jagetiya; 2020)

3rd-degree polynomial regression technique helped in predicting the daily growth rate of Covid-19 and required necessary measures to implement lockdown for the diminishing impact of Covid-19. A large number of people were still not tested, undetected and without implementing the lockdown measures, Covid-19 cases could be 6 times the cases as on 12th April 2020.

Artificial Intelligence (AI) is a powerful technology used in modern applications and has provided until now the best performance among other techniques. It helped in offering a quick and better approach in identifying the infected patients of Covid-19. Big data helped in supporting the models in forecasting the Covid-19 pandemic with the use of its ability to aggregate data in leveraging the huge amount of data to detect at an early stage. A data-driven strategy in learning the Covid-19 phenomenon to read the incubation period was proposed and analyzed it to be around 10 days (Pham et al.; 2020). Social networking platform played a vital role in better understanding and improvement of the pandemic situation. AI had been successful in finding the medical applications which aimed to restrict the disease transmission and focuses on the drug discovery and prediction of vitamin formation. It also helped in inventing new/existing medicines to be used for the immediate treatment of Covid-19 and also in case of no proper vaccination, it boosts the economy and benefits from the perspective of science. The models based on AI was observed to be suitable for the mitigation of the pandemic impact because of the availability of a huge amount of data and it was possible due to tremendous efforts and modern technologies. Big data played another crucial role in diminishing the spread of Covid-19 using data analytics to predict the future outbreak, especially on the international scope. Its ability to track the spreading of the virus was proposed to identify the pandemic areas which were at higher risks. Big data has the capability of supporting the treatment and diagnosis process of Covid-19. However there are some challenges which need to be addressed like changes in the government defining policies regularly, no defined dataset is publicly available and insecure data privacy and security. Further optimization using AI and big data was required to improve the performance and reliability to diagnose and treat in the best possible way.

Multiple machine learning algorithms were implemented on various Covid-19 dataset in last 6 months to predict or forecast the recovery rate, death rate or confirmed cases so that it might help people as well as the government authority to take necessary steps before it's too late. Therefore (Rustam et al.; 2020) implemented models such as support vector machine (SVM), linear regression, selection operator and exponential smoothing (ES) for the forecasting of new patients which might get affected due to Covid-19. The model (ES model) predicted for the duration of 66 days i.e. from 22/1/2020 to 27/3/2020 gave the best performance in forecasting the death rate, recovery rate and newly infected patients. The model performance was evaluated using R2 (R-squared), R2adjusted

(Adjusted R-square), MAE (mean absolute error), MSE (mean square error) and RMSE (root mean square error). LR model predicted death rates to increase in future days and a decrease in the recovery rate.
Another research (Jain et al.; 2020) predicted the growth in daily cases, daily positive cases and confirmed positive cases in India for 11 days using Generalized Growth Model (GGM), Generalized Logistic Growth Model (GLM) and Logistic Growth Model. High level of risk was estimated in India with approximately 12116 positive cases on 14th April and 25706 cases on 20th April. As growth rate was increasing exponentially, it was an indication of uncontrollable domestic transmission. Once again it was predicted that the national lockdown measures played a crucial role in curbing the surge in infected cases however, if the lockdown is lifted up without taking necessary precautions then the cases could rise. Due to improper reporting of actual cases and testing kits unavailability, predicted cases had a deviation with the actual cases and new hotspots origin led to inappropriate incident reporting in the month of April. The logistic model resulting in divergence showed the necessary actions to be taken as soon as possible in order to control the pandemic. There was an overlapping presence in the infected waves responsible for infection pools as defined by concatenated "Riccati module" (RM) (Marmarelis; 2020). Five different infected waves were detected in major infection pools using the analysis of daily Covid-19 data from 11th Mar to 18th June 2020 in the US. The 1st initial wave was smaller followed by 2nd one, which was larger, then 3rd wave corresponds to the urban surge and then 4th wave might spread across rural and smaller areas where the growth and size are lower and moderate, respectively. The 5th and the last infected wave were the largest wave detected on the 60th Day (9th May) and appeared to be responsible due to some relaxation in mitigation measures. The model predicted the total number of infected cases to be 4,160,000 which was actually the double of that time current combined number and new confirmed cases could drop to less than 5k by 20th September in the US. It was observed that the RM based technique outperformed simple curve-fitting algorithms due to is the ability of correlation with some alleviating measures.

3 Methodology

CRISP-DM (Cross Industry Process for Data Mining) methodology has been implemented in this research. This methodology provides structural pipeline and helps in robust data mining project planning. It is divided into six stages as displayed in figure 2:

3.1 Business Understanding

The world has been adversely impacted by Covid-19 resulting in loss of thousands of lives, millions infected and billions worth of money loss. This problem has been the motivation of this research project. Forecasting the number of cases and deaths using an accurate model can aid the governments in taking the precautionary measures on time which could eventually save as many lives and money possible.

3.2 Data Understanding

Dataset is the main aspect in the field of data analytics. Data collection and its understanding play a crucial role in the analysis and insights about the problem which needs to be solved. Data used for the research needs to be real and authentic. There are multiple

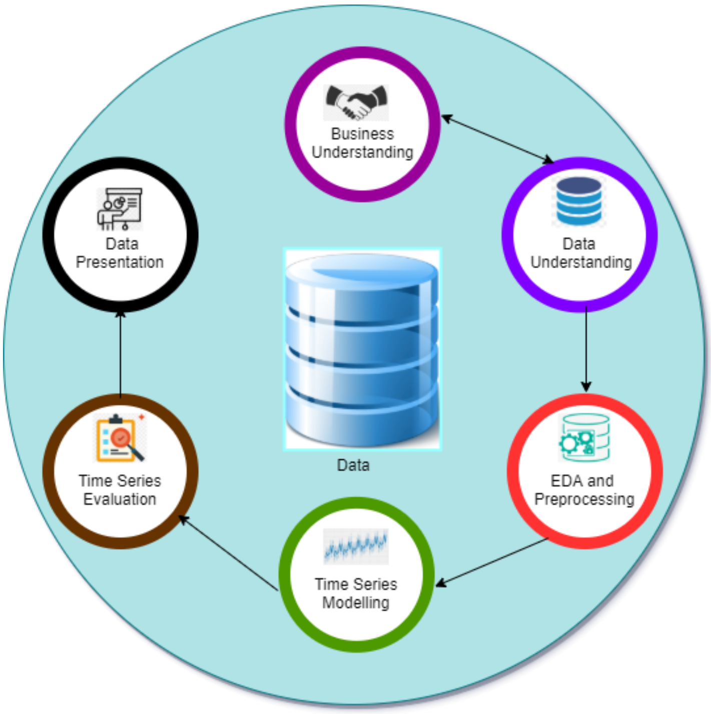


Figure 2: Modified Time Series CRISP-DM Methodology

sources where Covid-19 data is available and for this research, Covid-19 data was collected from Our World In Data[3] website. This website collects the data from the European CDC (Center for Disease Prevention and Control) website with the purpose of a global perspective. The dataset contains countrywise, continentwise and worldwide cases and deaths with many other related parameters. The dataset ranges from 31/12/2019 till 25/07/2020 which has daily data. New cases and new deaths for each day is added in the columns total cases and total deaths to get the cumulative number for each day. Dataset contains 32584 rows * 34 columns.

3.3 Exploratory Data Analysis and Data Pre-Processing

The dataset extracted was loaded into a data frame and then the structure of the dataset and missing values were analysed. All the columns were renamed into an easy, understandable and better feature names. Features with no importance for the analysis were dropped. Due to insufficient data of 'Hong Kong' and 'International' in-country observations, they were dropped and multiple data frames were created containing all countries data in one data frame, the cumulative sum of all countries data in one data frame and other continent wise data frames.

As shown in figure 3, variables such as new tests, total tests, new tests per thousand, total tests per thousand, new tests smoothed, new tests smoothed per thousand and test units have missing values more than 60%. After the analysis, it has been found that the government was not reporting the exact count of these variables and many countries did not have required test units to test all patients. There are only a couple of countries who have reported the test units from the starting of the pandemic.

Hence, Missing values have been filled with 0 because of the fact that there was no official

[3] https://ourworldindata.org/coronavirus-source-data

data available and imputing it with mean or median would adversely impact the analysis.

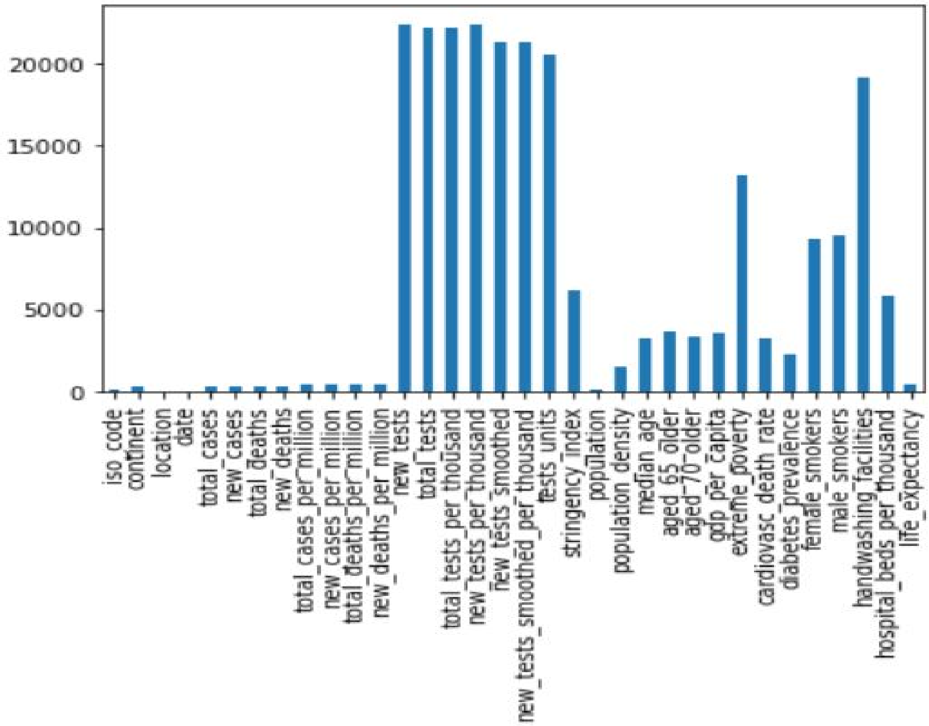


Figure 3: Missing Values

Correlation Analysis: As shown in the below figure 4, it can be seen that confirmed cases are strongly correlated with number of deaths and the population is also correlated with total cases and total deaths.

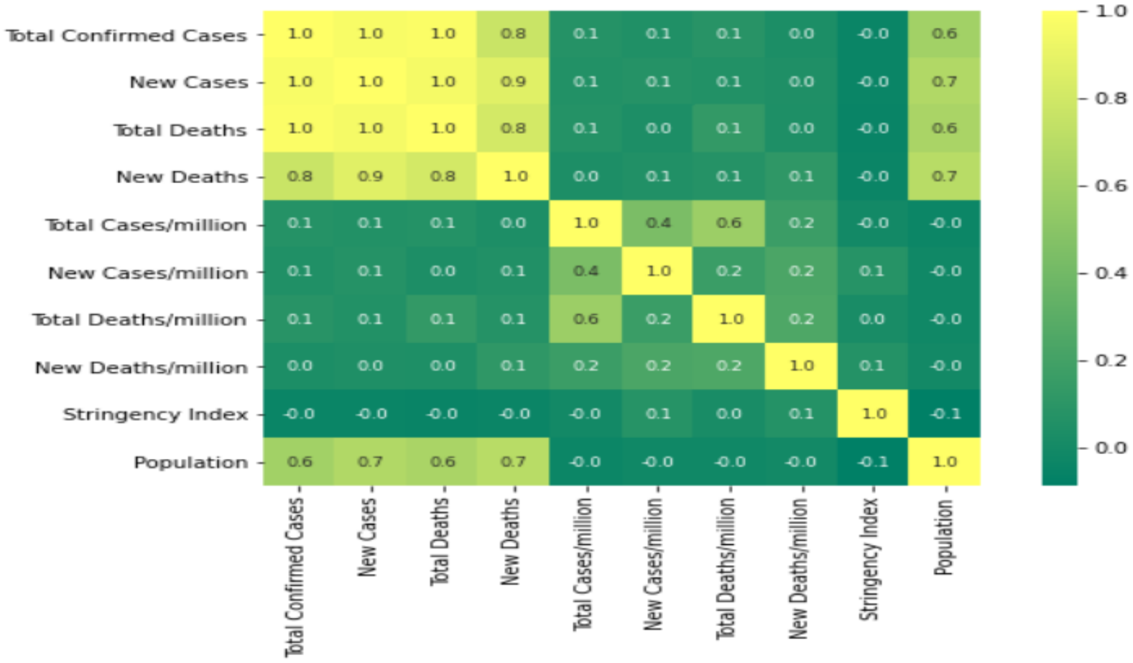


Figure 4: Correlation Analysis

As displayed in figure 5 and 6, the highest Covid-19 confirmed cases are reported in North America followed by Asia and the least cases in Oceania. However, total deaths reported in Europe and North America were almost the same followed by South America. There is speculation of European virus being more aggressive than Chinese, maybe due to the reason which is still not proved that China had been hiding the real confirmed cases and deaths. Also, at the starting of the pandemic, Italy had reported more cases than the beds in the hospital and all the European countries have reported the correct numbers from the starting. Although, China had imposed complete lockdown after Covid-19 was declared an international public health emergency.

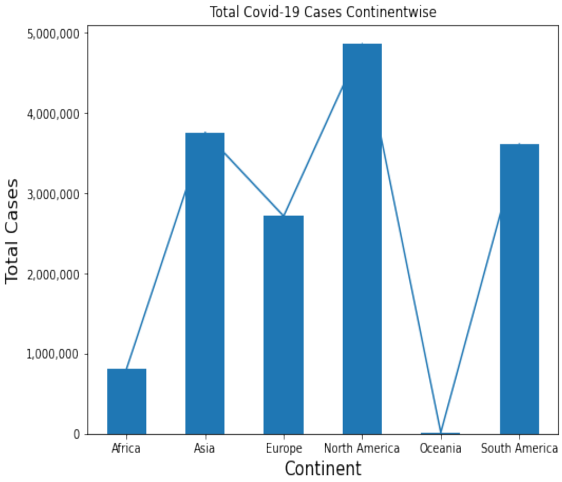


Figure 5: Total Cases

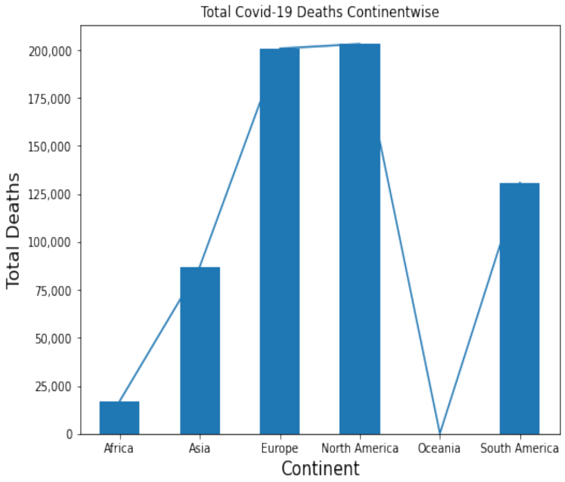


Figure 6: Total Deaths

Feature Engineering was performed to derive the new features out of the existing ones. Four new features were analysed as following:

- Date was transformed into Weekday - It will help in analyzing the number of days it took the cases and deaths to reach certain number
- Mortality Rate - It helps to know the percentage of infected people died due to Covid-19
- Growth factor was derived from dividing new cases by new cases shift by 1 observation - It is used to calculate the rate at which the number of cases is increasing per day
- New case peak to now ratio was derived as a ratio of a current new case and the max new case for each country

While extracting the data from the website, Spain's data for 25th July was not available, so the Exploratory data analysis was performed till date 24th July.

The United States is the most impacted country in the world with 4034102 cases as shown in figure 7. Brazil, India, Russia and South Africa are the other 4 countries which have been highly impacted by Covid-19. It is crucial to understand and analyze the impact of Covid-19 in the United States. United States were not prepared for such a rapid transmission of the disease and at the earlier stage of Covid-19, the country did not stop the flights coming from different parts of the world which might have eased the transmission of the disease.

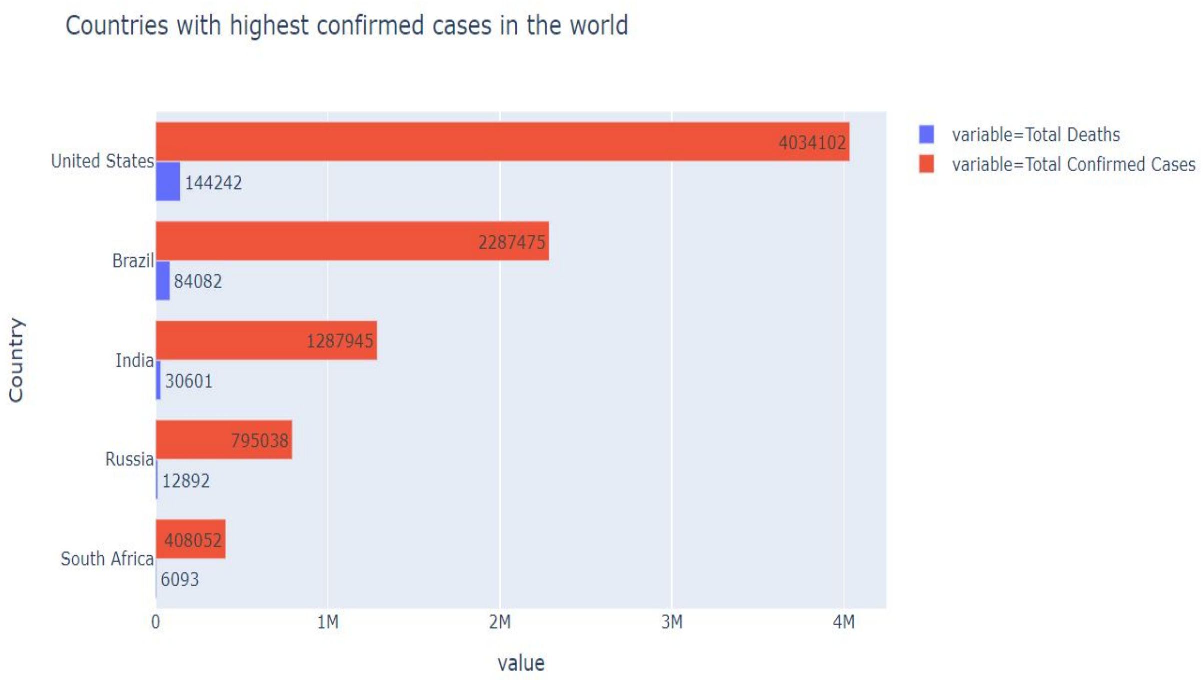


Figure 7: Countries with highest Covid-19 confirmed cases

People were not alarmed and no necessary precautions were taken by the government to tackle Covid-19. There was no proper lockdown in the country and people were refusing to wear masks which was strictly against the rules of WHO.

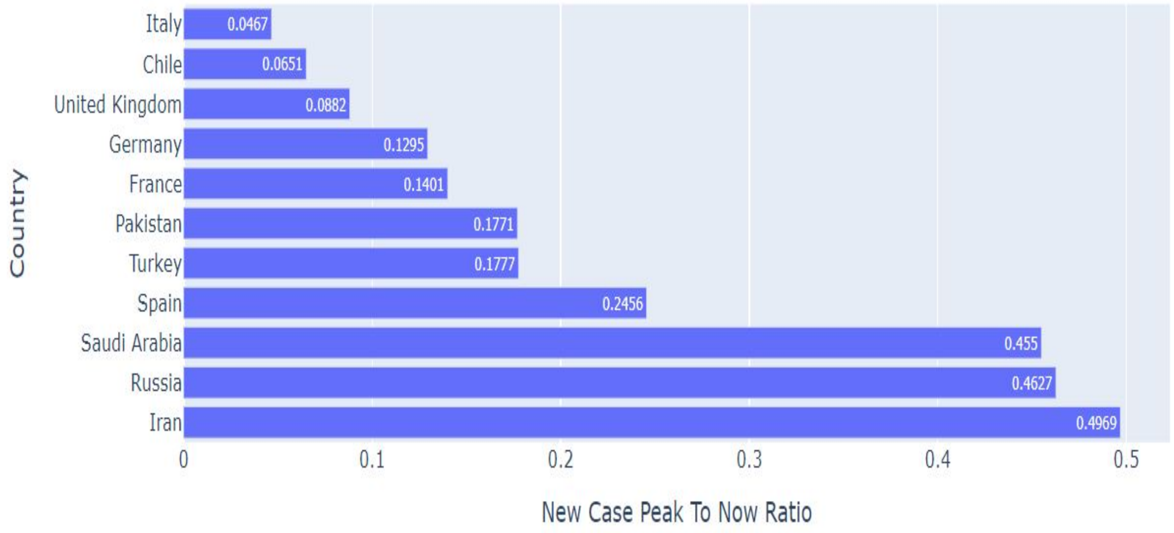


Figure 8: Countries with lowest mortality rate as on 25th July 2020

As displayed in above figure 8, when comparing countries with more than 150k confirmed cases, Italy has now the lowest mortality rate. In the month of March, Italy was the most affected country in the world. But after implementing lockdown measures, social distancing and mandatory wearing of masks, it brought down the infections to almost hundreds.

In the month of July, United States had been reporting more than 50k new cases per day while there is a fluctuation in the cases in Brazil. However, there is a exponential increase in the new cases reported per day in India starting from the month of May as shown in figure 9.

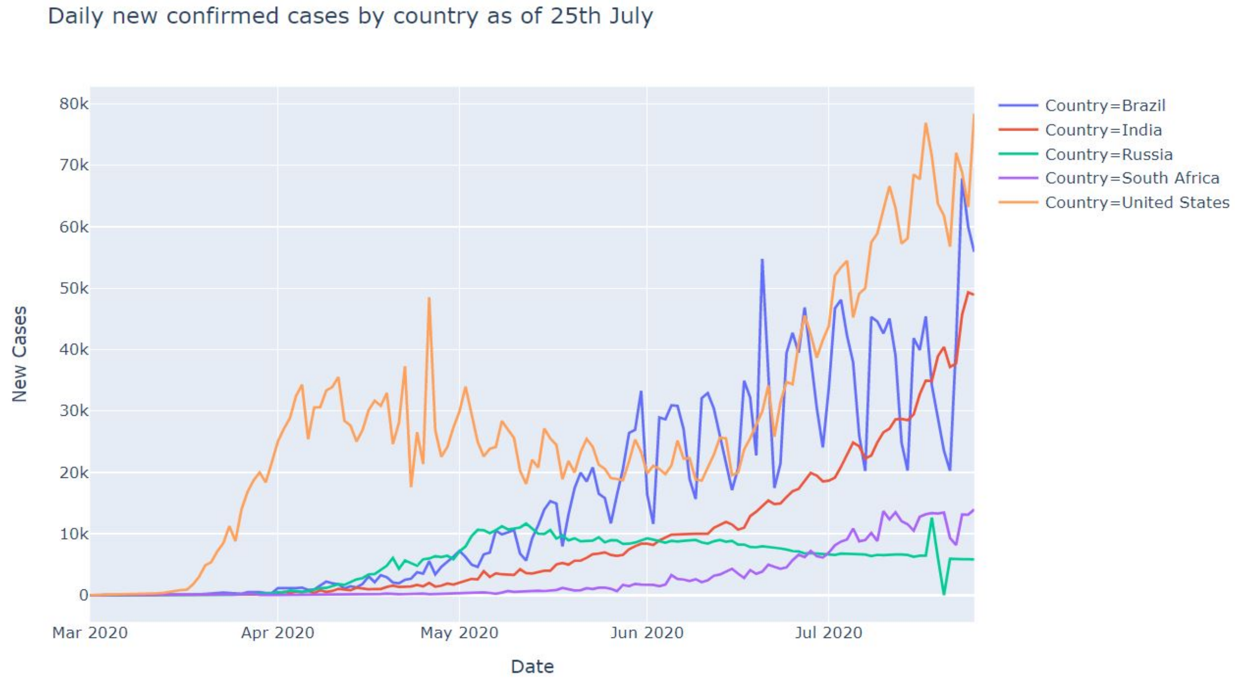


Figure 9: Top 5 Countries with highest daily new confirmed cases

Confirmed cases are increasing in India because of the lifting up of lockdown measure in various parts of the country and people not following the quarantine rules which is strictly applicable for all people either travelling from one state to another or any international arrivals.

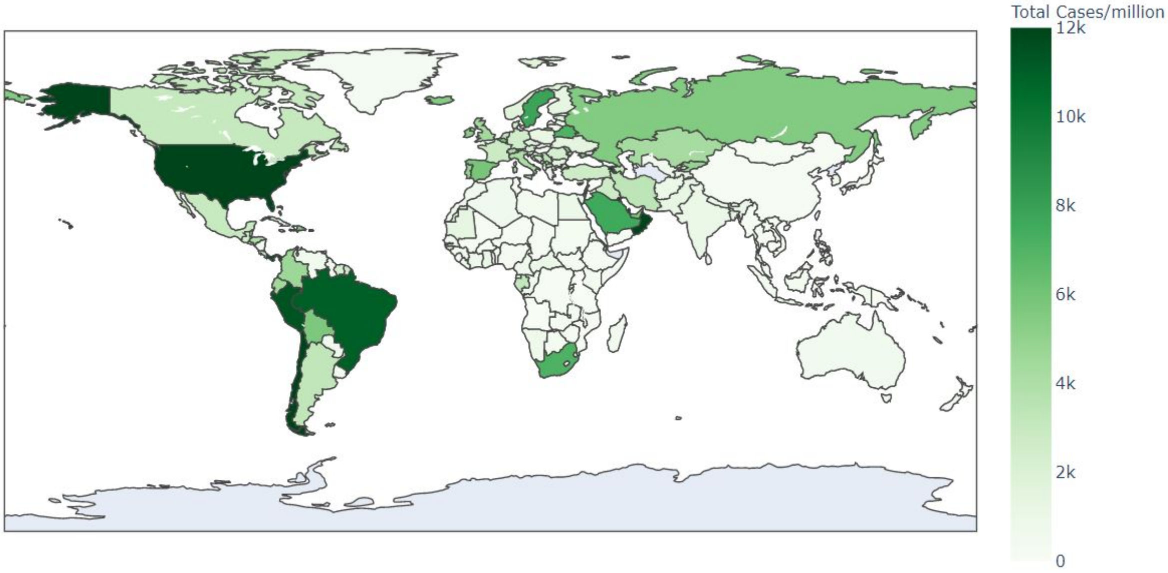


Figure 10: Worldwide Total Confirmed Cases per million

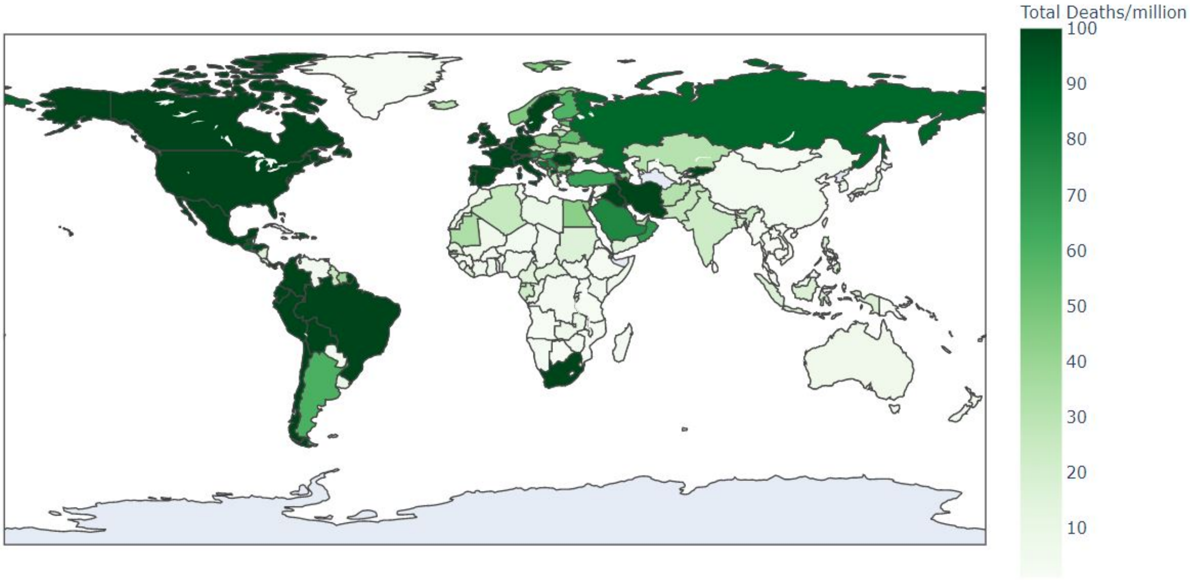
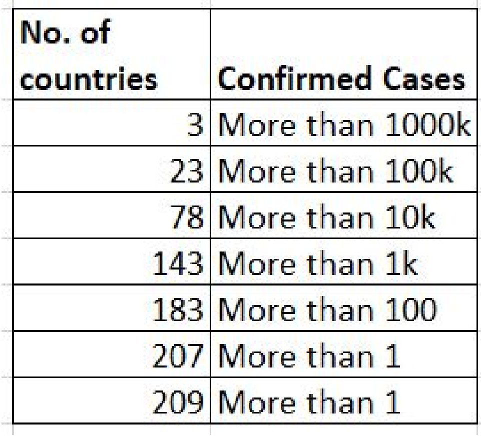


Figure 11: Worldwide Total Deaths per million

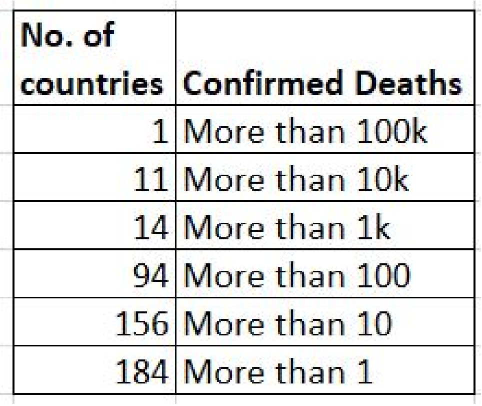
Figure 10 and 11 displays the worldwide confirmed cases and total deaths per million respectively as of 25th July 2020

Futhermore, the analysis on the count of countries impacted by Covid-19 is as displayed below:



No. of countries	Confirmed Cases
3	More than 1000k
23	More than 100k
78	More than 10k
143	More than 1k
183	More than 100
207	More than 1
209	More than 1

Table 1: Confirmed Cases



No. of countries	Confirmed Deaths
1	More than 100k
11	More than 10k
14	More than 1k
94	More than 100
156	More than 10
184	More than 1

Table 2: Confirmed Deaths

From the above tables, it can be observed that 209 countries have been impacted by Covid-19 and at least 1 death has been reported in 184 countries.

3.4 Time Series Forecasting Models

Some important tests are necessary to implement the models on the dataset before fitting the models. Dataset needs to be validated in order to be applied for the prediction and forecasting. In order to fit the models, total confirmed cases, total deaths and total new cases with the date and days have been used. Multiple models have been applied in this research for the prediction of total confirmed cases and total deaths due to Covid-19 and also forecasting for the next 59 days. Models applied in this research are as follows:

Polynomial Regression: This model provides better accuracy even if the relationship between predictor and target is correlated but not linear. It helps to fit in a broad function and curvature fits in it. It is implemented in a two-way process. Polynomial Regression with n degree is represented by:
Y = c_0 + c_1x + $c_2x^2 ... c_nx^n$
$where, c_0$ is the bias, c_1,c_2 and c_n are the weights in the equation and n is the polynomial degree.

Holt's Linear Model: This method is also known as linear exponential smoothing method popular for forecasting trend data. It works on three equation to produce the final forecast. It works on two parameters, first is the overall smoothing and second is the equation of the trend smoothing.

$$\begin{aligned} \text{Forecast equation} \quad & \hat{y}_{t+h|t} = \ell_t + hb_t \\ \text{Level equation} \quad & \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\ \text{Trend equation} \quad & b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}, \end{aligned}$$

Auto ARIMA-ARIMA Model: ARIMA (Auto Regressive Integrated Moving Average) model is used for time series prediction and forecasting. It uses previous logged time series observations to forecast the new observations. Auto Regressive (p) and Moving Average (q) are the two major components in ARIMA model. This model involves three steps as follows:
1.**Identification:** This step is used in finding the actual values of auto regressive (p), number of differencing (d) and moving average(q). Values of p and q can be identified by partial correlation function and autocorrelation function.
2.**Estimation:** After picking the correct ARIMA (p,d,q), model can be fit and predicted.
3.**Evaluation:** The model's performance is evaluated using metrics such as RMSE, MAE, AIC and BIC value.

Auto ARIMA-AR Model: AR (Auto Regressive) also known as AR(p) model is a time series model which predicts the next value from using historical observations as an regression input equation. It is defined as:
y_t=b_0+$b_1$$y_{t-1}$+$c_t$
where, y_t = time series value at time t
b_0 = intercept at the vertical axis (y)
b_1 = coefficient of the slope y_{t-1} = value of time series at time t-1 c_t = error term

Prophet Model: A novel forecasting model developed by Facebook which can deal with hourly, daily or weekly observations and has the capability to handle time-series features like outliers, trend, holiday and seasonality[4]. To use the prophet model, 'Prophet' function is imported from 'fbprophet' library. Predictor and the target are converted into columns 'ds' and 'y' for prediction. The predictions are then saved in a future data frame which contains the predicted dates as an index 'make future data frame' function was used to create the data frame and 'predict' was used to forecast values. Weekly and trend seasonality were plotted using an inbuilt prophet feature 'plot components'. Prophet's inbuilt cross-validation was also performed.

[4] https://research.fb.com/blog/2017/02/prophet-forecasting-at-scale/

3.5 Time Series Evaluation

Evaluation metrics has been carried out on the applied models to check the results achieved. All the models have been evaluated and forecast for the next 59 days. Following are the metrics used in this research to measure the model's performance:

Mean Absolute Error: MAE is calculated by taking the average of all forecast errors where only positive values of forecasting are considered. MAE uses the difference of test observations with that of predicted absolute forecast values and is also not sensitive to outliers. This is the reason MAE outperforms mean squared error as it is outlier sensitive. it is considered as the primary performance metrics due to multiple peak values in the data.

Root Mean Squared Error: RMSE is the most frequent evaluation metrics used in time series models which measures the differences of the model predicted values with the observed values. It tells us the concentration of the data around the best line fit.

Mean Absolute Percentage Error: MAPE is scale-independent. Metrics like MSE and RMSE are dependent on scaling and not reliable on different data scaling. As there is only one predictor or its a univariate time series, MAPE can provide better performance evaluation. It gives the average performance errors of forecasts.

3.6 Data Presentation

Being an academic project, the project will be presented with all the aspects of work performed as well as the result and conclusion in the form of video presentation although this stage is commonly referred to as Deployment.

4 Implementation

This section provides a detailed description of the models implemented to predict and forecast the total confirmed cases and deaths as listed in the above research objective. Also, the performance of all the models will be discussed. We will discuss the best 3 models as per the evaluation metrics and the remaining models will be discussed in the configuration manual.

4.1 Prediction and Forecasting of Covid-19 total confirmed cases

AR-ARIMA Model: The model was trained on p-value ranging from 0 to 5 with differencing as 2 and the lowest Akaike Information Criteria (AIC) and Bayesian information criterion (BIC) was found for AR Model (5,2,0). The model was predicted with no seasonality.

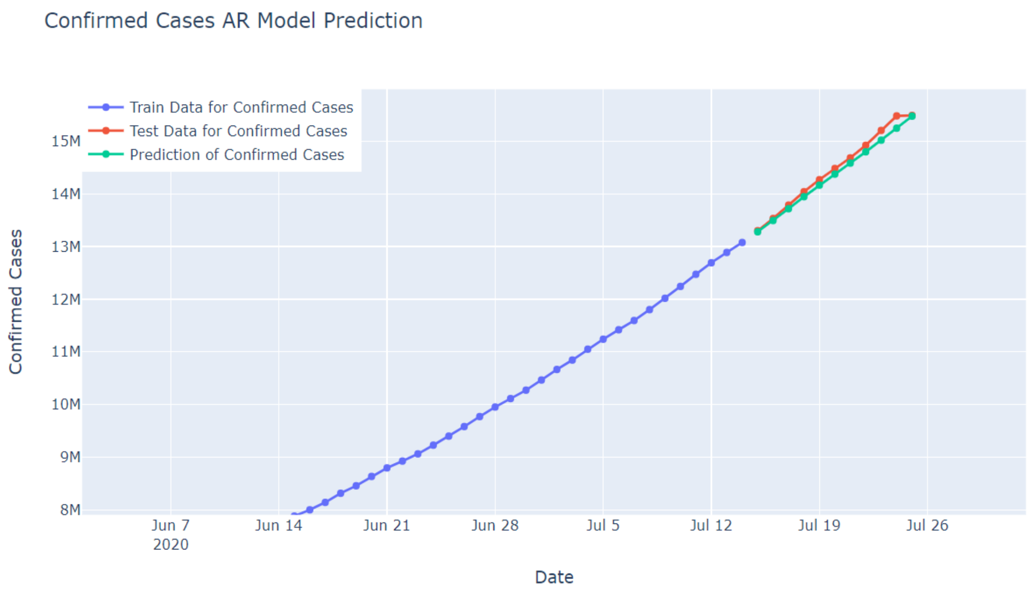


Figure 12: ARIMA-AR Model Prediction - Covid-19 Confirmed Cases

From the closer look of the above figure 12, we can see that the prediction of total confirmed cases by AR-model on the test set is good and almost near to the original data.

PROPHET Model: With the implementation of Prophet Model, better insights about the peak and fall of the data was achieved as compared to other models.
From the below figure 13, we can see that prediction of total confirmed cases by Prophet-Model on the test set is better than AR-model but not ARIMA-Model. Here, y (target) is the total confirmed cases and ds (predictor) is the Date.
Also, by using components feature as shown in figure 14, Prophet model found more confirmed cases to be reported on weekends and trend seemed to be exponentially increasing for the next forecast 59 days. **(Please Note : le7 = 10million)**

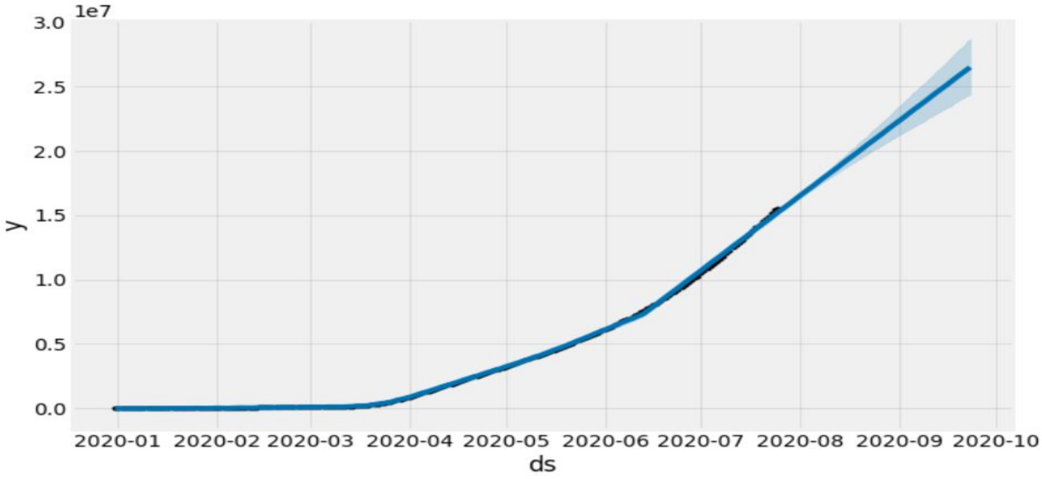


Figure 13: Prophet Model Prediction and Forecasting

The blue line shows the original confirmed cases and blue lines are the predicted and forecast confirmed cases.

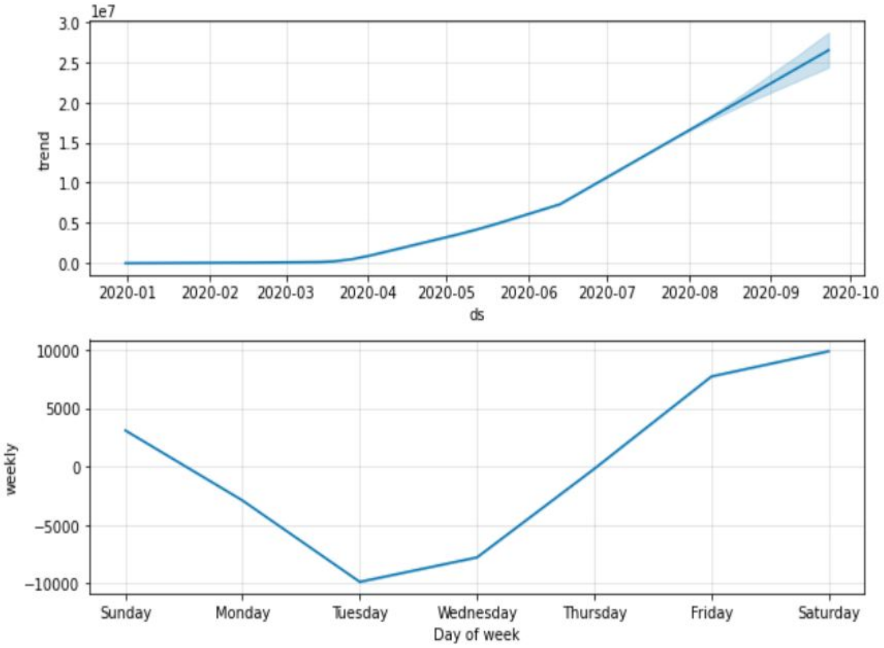


Figure 14: Prophet Model Components

Maximum cases forecasted are reported in the weekends, which might be due to the reason that people go out in weekdays and then gets infected and take the test. As the test results take 2-3 days, it is expected to be reported on weekends.

ARIMA Model: The model was trained on p and q value ranging from 0 to 3 each and differencing as 2 and the lowest Akaike Information Criteria (AIC) and Bayesian information criterion (BIC) was found for AR Model (3,2,2). This model was predicted with no seasonality.

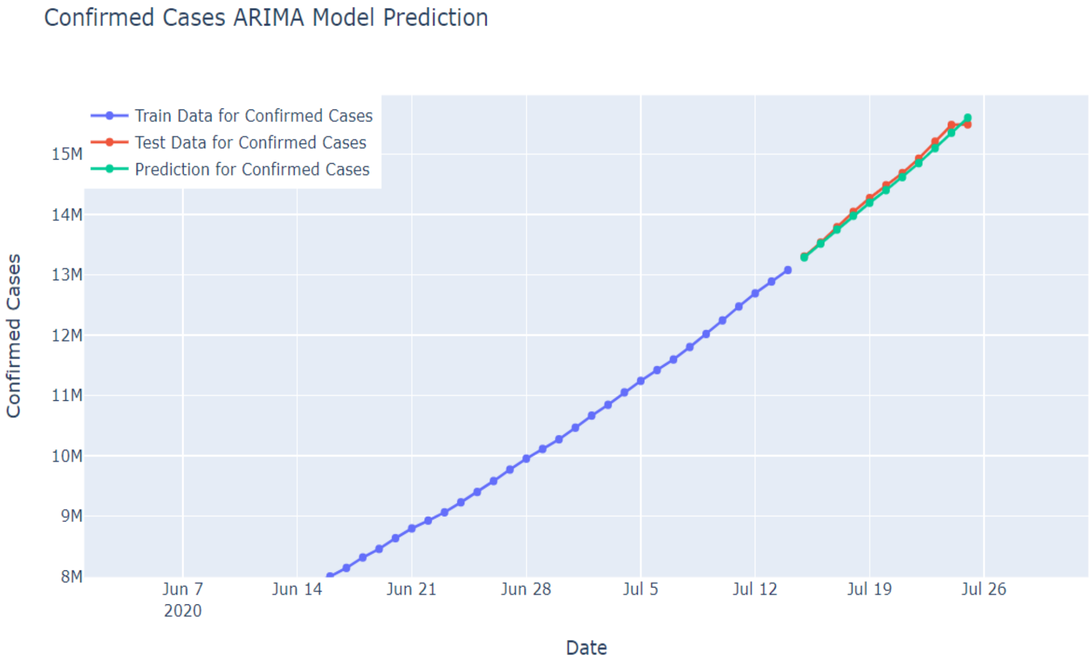
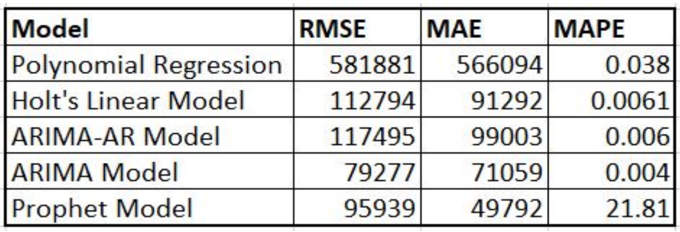


Figure 15: ARIMA Model Prediction - Covid-19 Confirmed Cases

4.1.1 Evaluation Analysis:

The performance of all models are tabulated as shown in table 3.



Model	RMSE	MAE	MAPE
Polynomial Regression	581881	566094	0.038
Holt's Linear Model	112794	91292	0.0061
ARIMA-AR Model	117495	99003	0.006
ARIMA Model	79277	71059	0.004
Prophet Model	95939	49792	21.81

Table 3: Performance Comparison

It is clearly evident from the above table, the ARIMA Model performed the best with achieving the lowest error in all evaluation metrics. ARIMA Model outperformed all other models in predicting Covid-19 total confirmed cases.
ARIMA model forecasts confirmed cases to go past 26 million cases by 31st August 2020 while other models forecast fewer cases as shown in figure 16.
The live Covid-19 data for the comparison can be found at the following website:
`https://www.worldometers.info/coronavirus/`
Also, if we look at figure 16, it can be seen that there is a big difference between the total confirmed cases forecasted by 5 models. Forecasting has been displayed from 17th to 31st Aug 2020.

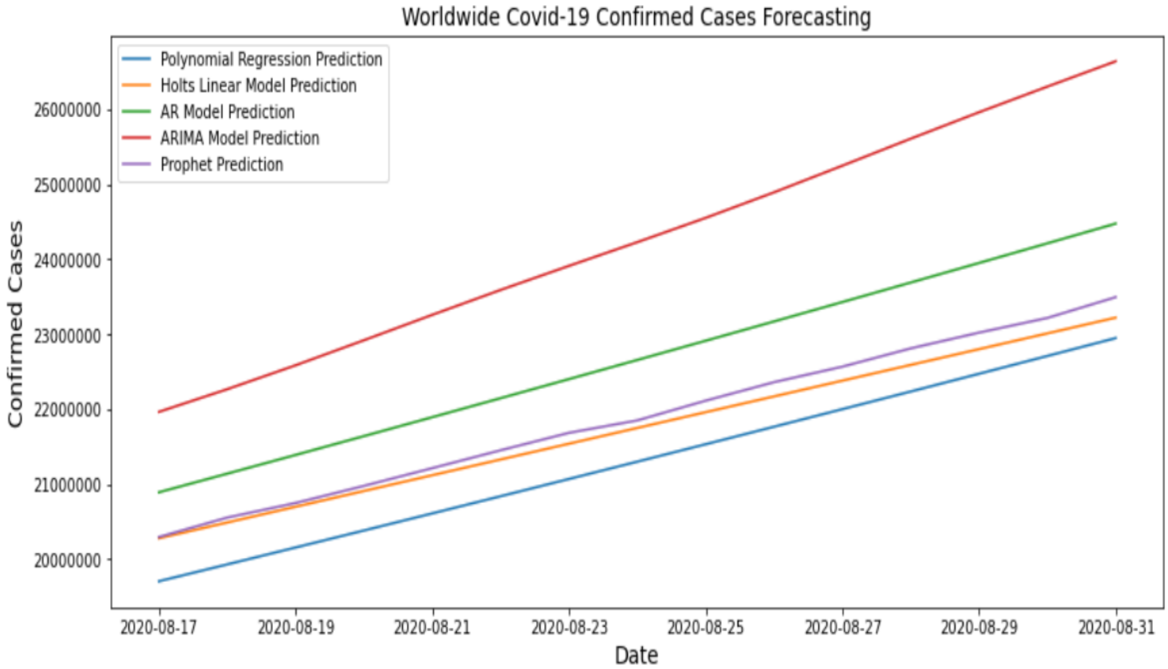


Figure 16: Worldwide Confirmed Cases Forecasting

4.2 Prediction and Forecasting of Covid-19 total deaths

Now, we will predict and forecast the worldwide total deaths due to Covid-19. As we know the ARIMA performed the best in confirmed cases prediction, we will compare it with Prophet model for total deaths forecasting.

ARIMA Model: The model was trained on p and q value ranging from 1 to 3 each and differencing as 2 and the lowest Akaike Information Criteria (AIC) and Bayesian information criterion (BIC) was found for AR Model (2,2,3). Again, the model was predicted with no seasonality.

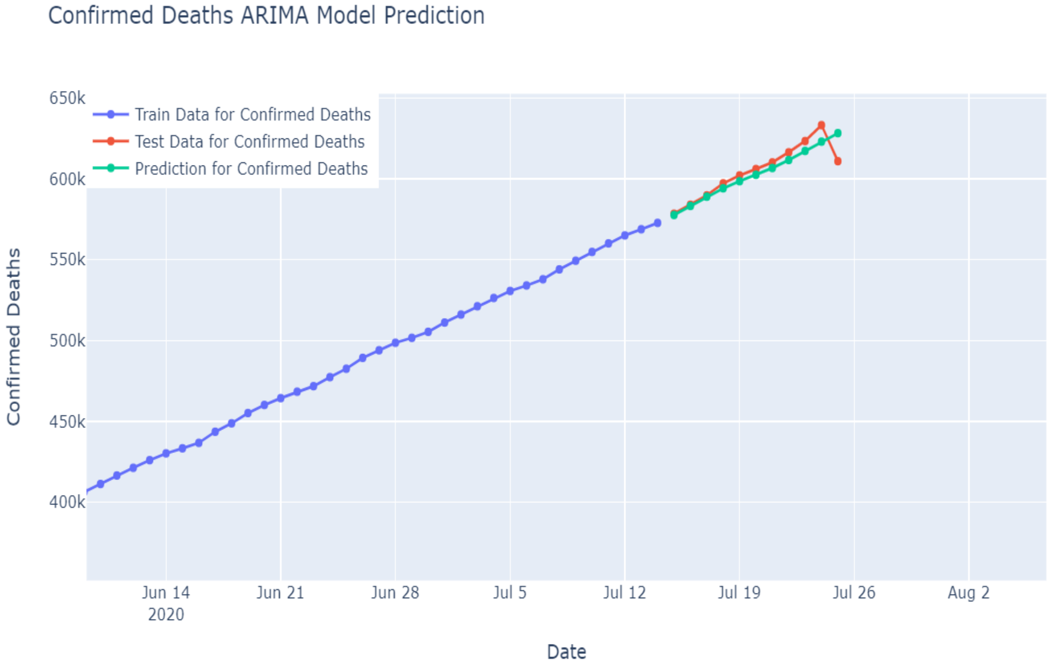


Figure 17: ARIMA Model Prediction - Covid-19 Confirmed Deaths

Prophet Model: Using the same criteria used for confirmed cases above, Prophet model has been implemented to predict and forecast the total confirmed deaths.

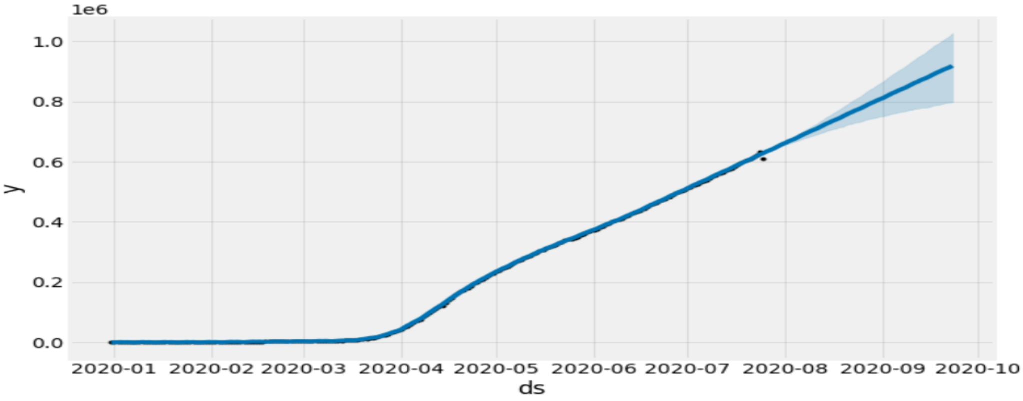


Figure 18: Prophet Prediction and Forecasting - Covid-19 Confirmed Deaths

From the above figures 18, it is clear that the prophet model has predicted the confirmed deaths which is almost close to the original deaths and most of the deaths have been forecasted to be on weekends mainly Friday as shown in figure 19. Also, we found the highest cases to be on Friday (Prophet model) so we can conclude that timespan between people getting infected and dying is either the same day or 7 or 14 days.

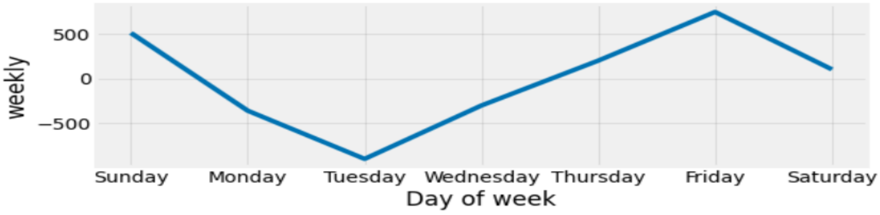
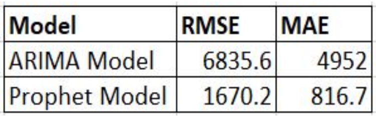


Figure 19: Prophet Model Components

4.2.1 Evaluation Analysis:

Below table 4 provides the performance of the models applied and we find that the Prophet model outperformed the ARIMA model. It acquired the lowest errors in all evaluation metrics.



Model	RMSE	MAE
ARIMA Model	6835.6	4952
Prophet Model	1670.2	816.7

Table 4: Performance Comparison

From the below model forecasting as displayed in figure 20, it can be seen that the total deaths will exponential increase till 31st August although Prophet model forecasts 860k deaths and ARIMA model forecasts 800k respectively by 31st Aug 2020.

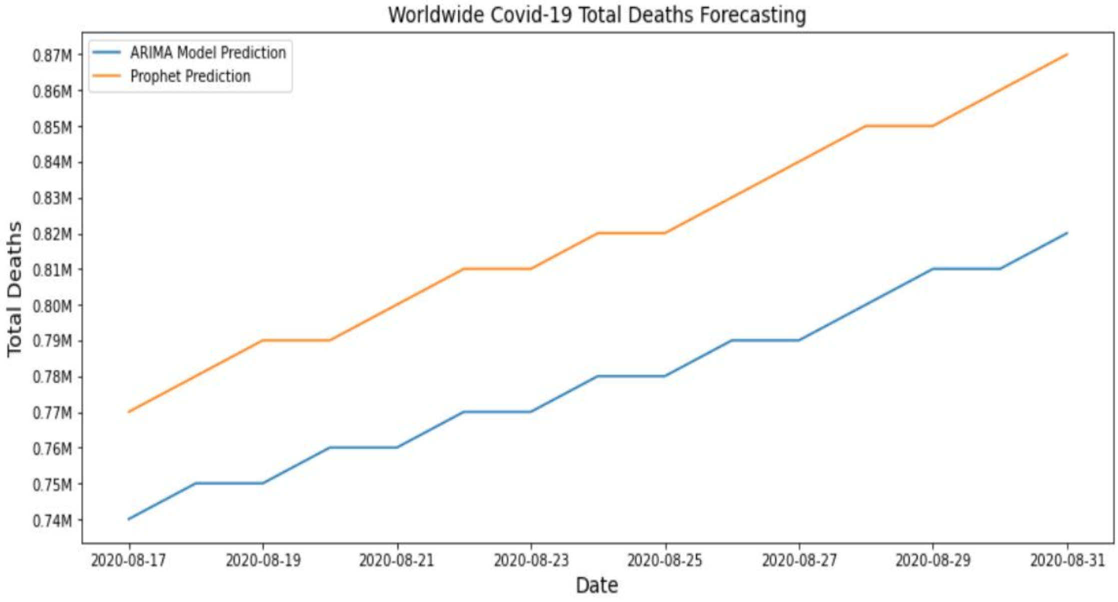


Figure 20: Worldwide Total Deaths Forecasting

5 Discussion

The research implemented in this paper provides an insight into the Covid-19 which created a catastrophic environment all over the world. There is an utmost need for forecasting the total cases and deaths due to Covid-19 so that the governments of all the countries can take highly necessary precautionary measure. Also, it would help healthcare professionals to identify the cause of increasing infections and provide their advice to the people on taking necessary steps to lower the reproduction number. The dataset used in this research contained daily worldwide data from 31/12/19 till 25/07/20. Exploratory data analysis and pre-processing found the cases and deaths to be exponentially increasing over the time.

ARIMA Model outperformed other models to predict and forecast the total confirmed cases with RMSE: 79277, MAE: 71059 and MAPE: 0.004. Prophet model found a higher number of cases to be reported on the weekends. Also, at the time of writing this report (06/07/2020), it is observed that the actual confirmed cases(18.8M) and forecast confirmed cases(18.7M) by ARIMA Model have less difference. Similarly, ARIMA and Prophet Models were implemented to predict and forecast total deaths due to Covid-19 and it was evaluated that the performance of Prophet model outperformed ARIMA with RMSE: 1670 and MAE: 816.7. Also, the higher number of deaths is forecasted to be on the weekends.

6 Conclusion and Future Work

The primary objective of this research paper was to alert the government and public by providing them with the forecasting number of infections and deaths due to Covid-19. It also provides the brief details of the cases and deaths reported in each continent and the highest mortality rate in different countries. The models implemented gives acceptable and almost close to the original data prediction results with reference to ARIMA, AR-ARIMA and PROPHET model. Therefore, it can be illustrated that the machine learning approach can be useful for Covid-19 cases and deaths forecasting.

There is a strong correlation among total cases, total deaths and total tests performed. Also, it can be seen that the Covid-19 cases and deaths will exponentially increase in the forecasted next 59 days. Most numbers of cases are reported in weekends due to the reason that people come out of their home in weekdays either for work or personal stuff and then they get infected, take the tests and find out the result which takes 2-3 days.

There is an urgent need to stop spreading this deadly virus and therefore, the government has to educate people on taking the necessary precautions, people should wear masks in all public places and there is an urgent need to test people as many possible to know the number of people infected and putting them in quarantine. Also, the forecast high numbers is a great health concern for every citizen and there is a need for more lockdowns in future to completely stop this virus spread. Hospitals can estimate the number of patients to be admitted and the count of beds and ventilators to be required from this research analysis. However, there is an uncertainty in the people getting infected which could deviate the forecasting data. So in future, continuous forecasting of the Covid-19 cases and deaths is required because of the reason that may be lockdown is lifted up and infections increases or if there is a complete lockdown then virus stops spreading.

7 Acknowledgement

I would like to thank my mentor, Prof. Christian Horn, for his patience and encouragement throughout the semester. He provided the mental and moral support for my project. His assistance helped me to achieve tremendous knowledge about machine learning techniques and other related work in this research. I would also like to thank the authors of Our World in Data website for providing a structured dataset which was used in this research.

References

Bhati, A. and Jagetiya, A. (2020). Prediction of covid-19 outbreak in india adopting bhilwara model of containment, *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pp. 951–956.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y. et al. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study, *The Lancet* **395**(10223): 507–513.

Fanelli, D. and Piazza, F. (2020). Analysis and forecast of covid-19 spreading in china, italy and france, *Chaos, Solitons Fractals* **134**: 109761.

Farahi, Z. and Kamandi, A. (2020). Coronavirus spreading analysis using dynamic spreading factor epidemic models, *2020 6th International Conference on Web Research (ICWR)*, pp. 184–190.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J. and Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in wuhan, china, *The Lancet* **395**(10223): 497 – 506.

Jain, M., Bhati, P. K., Kataria, P. and Kumar, R. (2020). Modelling logistic growth model for covid-19 pandemic in india, *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pp. 784–789.

Kolla, B. (2020). Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms, *International Journal of Emerging Trends in Engineering Research* **8**.

Lai, C.-C., Shih, T.-P., Ko, W.-C., Tang, H.-J. and Hsueh, P.-R. (2020). Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and coronavirus disease-2019 (covid-19): The epidemic and the challenges, *International Journal of Antimicrobial Agents* **55**(3): 105924.

Mahase, E. (2020). China coronavirus: Who declares international emergency as death toll exceeds 200, *BMJ: British Medical Journal (Online)* **368**.

Mandal, M., Jana, S., Nandi, S. K., Khatua, A., Adak, S. and Kar, T. (2020). A model based study on the dynamics of covid-19: Prediction and control, *Chaos, Solitons & Fractals* p. 109889.

Marmarelis, V. (2020). Predictive modeling of covid-19 data in the us: Adaptive phase-space approach, *IEEE Open Journal of Engineering in Medicine and Biology* pp. 1–1.

Nishiura, H., Jung, S.-m., Linton, N. M., Kinoshita, R., Yang, Y., Hayashi, K., Kobayashi, T., Yuan, B. and Akhmetzhanov, A. R. (2020). The extent of transmission of novel coronavirus in wuhan, china, 2020.

Organization, W. H., health organization, W. et al. (2020). Coronavirus disease (covid-2019) situation reports.

Pham, Q., Nguyen, D. C., Huynh-The, T., Hwang, W. and Pathirana, P. N. (2020). Artificial intelligence (ai) and big data for coronavirus (covid-19) pandemic: A survey on the state-of-the-arts, *IEEE Access* pp. 1–1.

Remuzzi, A. and Remuzzi, G. (2020). Covid-19 and italy: what next?, *The Lancet* **395**(10231): 1225 – 1228.

Roda, W. C., Varughese, M. B., Han, D. and Li, M. Y. (2020). Why is it difficult to accurately predict the covid-19 epidemic?, *Infectious Disease Modelling* **5**: 271 – 281.

Rothan, H. A. and Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak, *Journal of Autoimmunity* **109**: 102433.

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B., Aslam, W. and Choi, G. S. (2020). Covid-19 future forecasting using supervised machine learning models, *IEEE Access* **8**: 101489–101499.

Wang, J. and Du, G. (2020). Covid-19 may transmit through aerosol, *Irish Journal of Medical Science (1971-)* pp. 1–2.

Wang, L., Li, J., Guo, S., Xie, N., Yao, L., Cao, Y., Day, S. W., Howard, S. C., Graff, J. C., Gu, T., Ji, J., Gu, W. and Sun, D. (2020). Real-time estimation and prediction of mortality caused by covid-19 with patient information based algorithm, *Science of The Total Environment* **727**: 138394.

Wang, W., Tang, J. and Wei, F. (2020). Updated understanding of the outbreak of 2019 novel coronavirus (2019-ncov) in wuhan, china, *Journal of Medical Virology* **92**(4): 441–447.

Yang, J., Zheng, Y., Gou, X., Pu, K., Chen, Z., Guo, Q., Ji, R., Wang, H., Wang, Y. and Zhou, Y. (2020). Prevalence of comorbidities and its effects in patients infected with sars-cov-2: a systematic review and meta-analysis, *International Journal of Infectious Diseases* **94**: 91 – 95.

Zheng, N., Du, S., Wang, J., Zhang, H., Cui, W., Kang, Z., Yang, T., Lou, B., Chi, Y., Long, H., Ma, M., Yuan, Q., Zhang, S., Zhang, D., Ye, F. and Xin, J. (2020). Predicting covid-19 in china using hybrid ai model, *IEEE Transactions on Cybernetics* **50**(7): 2891–2904.

Zhong, L., Mu, L., Li, J., Wang, J., Yin, Z. and Liu, D. (2020). Early prediction of the 2019 novel coronavirus outbreak in the mainland china based on simple mathematical model, *IEEE Access* **8**: 51761–51769.