Case Studies in Data Science (COSC2669)

Final Report
WIL Project
COVID19 CHALLENGES

Group49

Shubhankar Jahagirdar, Pranamya Korde ,

Gaurav Sinha, Tanmay Shendkar.

RMIT University

Shubhankar (s3793593@student.rmit.edu.au)

Pranamya (s3779009 @student.rmit.edu.au)

Gaurav (s3807633 @student.rmit.edu.au)

Tanmay (s3735580@student.rmit.edu.au)

21st October 2020

Contents

Team Member	% Contribution
Shubhankar (s3793593@student.rmit.edu.au)	**25**
Pranamya (s3779009@student.rmit.edu.au)	**25**
Gaurav (s3807633 @student.rmit.edu.au))	**25**
Tanmay (s3735580@student.rmit.edu.au)	**25**
Total	***100***

1.0 Abstract

COVID-19 is a pandemic that has infected more than 170 nations worldwide. In almost all the affected nations, the number of infected and deceased patients has been rising at an alarming pace.

This report highlights the importance of country lockdown and self-isolation InControl the disease transmissibility among population throughout the world.

It is possible to inculcate forecasting methods to help design better plans and to make productive decisions. These methods analyse the circumstances of the past, thereby allowing better assumptions about the situation to take place in the future. Such forecasts may help plan for future threats and consequences. In providing reliable forecasts, forecasting techniques play a very important role.

Data collected from various platforms also play a vital role in forecasting. Data we used for our forecasting was from John Hopkins Website, WHO.

Researchers are working to investigate efficient and accurate models in order to predict the death count. Researchers are also working to provide a list of guidelines that can be followed by the people to reduce the spread rate of the COVID-19.

Our study indicates that there is a need to reassess control measures,

initiated by countries. Prediction of the spread and reproduction number should be analysed on varied datasets.

Forecasting methods, however, come with their own range of challenges. This report explores these issues and also offers a set of guidelines for individuals currently battling the global pandemic of COVID-19.

Forecasting and proper study of the pattern of disease spread could be very helpful in the planning of control strategies.

2.0 Introduction

The latest destructive pandemic, COVID-19, is running its course around the world. Not only are economies collapsing, but the overall strengths and principles of the nations that have been significantly affected are being undermined. On March 11, the World Health Organization (WHO) announced a "global pandemic" of a new pneumonia epidemic. Due to its tremendous spreading ability and possible damage, the new coronavirus has caused a great danger to the health and safety of people around the world. One thing has been found throughout the history of these epidemics, that is, with the advancement of time, these epidemics escalated into pandemics, or sometimes referred to as the virus / disease outbreak. A disease escalates into a pandemic when the situation at the local source where the outbreak was first observed to spread gets out of control. To reduce the transmission of COVID-19, many countries had instituted large-scale or national closure of schools by March, 2020. These actions appear largely based on assumptions that the benefits apparent in influenza outbreaks are also likely to be true for COVID-19. There are several theoretical reasons why school closures might be less effective in COVID-19 than in influenza outbreaks.
To date, 181 (by 5 April) countries with more than 1100,000 confirmed cases have been affected by the coronavirus and about 65,000 people have lost their lives. The 2019-2020 outbreak of COVID-19 is now officially recognised as a pandemic by the WHO, with the United States, Spain, Italy, and Germany suffering the worst cases of outbreaks and showing no sign of alleviation. An outbreak or epidemic also refers, at a specific time and location, to a sudden rise in the incidence of infectious diseases. Pandemics are near-global disease outbreaks affecting numerous nations around the world. What is the main difference of COVID-19 outbreak with pervious epidemics? Fortunately, the daily case detection changes are available and can be tracked almost in real time with predicting techniques such as forecasting.
Whenever these pandemics occur, world economies are majorly hit. Billions of dollars need to be invested in controlling an outbreak as well as in the development of a vaccine for the new disease. The main aim of our study is the analysis of forecasting techniques in computing and processing perspective. Forecasting will help to make reliable prediction and estimates with less computational overhead and there is a lack of bias.
Forecasting and proper study of the pattern of disease spread could be very helpful in the planning of control strategies. At this stage, a complete lockdown imposed in the affected area (already implemented by many countries) is a good solution to prevent and hopefully stop the spread (local transmission). We hope that by providing analysis of various forecasting models of COVID-19 will be more helpful for adapting better intervention policies and explicitly, it will also help to alleviate the alarming effect of this pandemic

Every forecast is carried out with some perspective irrespective of which category it may represent. From these studies and the forecasts made, it is very clear that the major outcome is to support healthcare communities to initiate critical action, decisions, control measures and public restrictions in time. Another outcome is to support in establishing mechanisms that provide control measures to be considered internationally for the global control of this pandemic as well as restrictions to the public in terms of quarantine, isolation, contact tracing, the recommendation in terms of metrological conditions (mainly Air, Temperature, relative humidity, wind speed and visibility) and its impact on the spread.

3.0 Project Aim

3.1 Problem Statement

As predicted, there are several obstacles for medical and governmental authorities to efficiently manage this respiratory illness. In spite of appropriated supplies, most hospitals are suffering from a scarcity of free beds, protective masks, sanitizing liquids and even ECMO machines for patients with severe cases. Defeating this pandemic is impossible without united and coordinated international attempts shaped by all countries of the world. We believe that an international scaled-determination is required to diminish the complex impacts of pandemic. The most important priorities are supposed to be:

i) The development of potential vaccine candidates to provide protection and interrupt the transmission of SARS-CoV-2

ii) To ensure enough supplies for hospitals and their homogeneous distribution among the countries with the worst number of severe cases.

iii) There is a need for more studies to identify potential treatments that are effective for the control of this viral infection

iv) It is imperative to provide easy access to diagnostic kits for all countries affected by this pandemic.

In the light of these suggestions, it would be recommendable to at least temporarily abandon the political checkouts in both national and international levels; therefore, all partners will be potentially able to efficiently enforce their strategies for the elimination of this unique threat to the human populations.

3.2 Project Objective

As discussed in the above section, the problem and the concern raised by COVID19 has really taken the world to critically think about all critical sectors which became non-functional with this pandemic.

The objectives of the study are as follows:

1. To study existing forecasting models.
2. To categorize forecasting models based on type of datasets.
3. To study of symptomatic and asymptomatic parameters.
4. To derive challenges related to forecasting models.
5. To formulate recommendations to control the pandemic.

This study is organized into four main sections. The experiment starts with the natural course of the disease; categorization of the diseases, along with the global history of pandemics where the COVID-19 outbreak is also mentioned. In the "Experiments" section - Project provides insights to the Experiments conducted on the data to evaluate the correct COVID 19 trends of confirmed cases for next 15 days . The implementation of various Machine Learning models which fits the correct data structure will be shared in the "Results" Section. The Limitations regarding the experiments and inconsistencies observed in experiments and results will be discussed in "Limitations" and finally the success story of predicting the 15 days of COVID 19 confirmed cases will be shared in "Conclusion" section.

3.3 Proposed Solution

This project Aim is to provide a solution to all the problem statements mentioned above by creating a one stop shop solution - an active dashboard with multiple tabs which holds all required and authentic data trends for latest information on COVID 19.

❖ **Delivered an Active Dashboard with Data and Graphs:**

- *Dashboard < COVID Country Trends >*
  - All Data and Trends of COVID (world wide)
  - Individual Country Stats from 22 Jan - Present
  - Active Bubble chart to Showcase Major impacted World Region
  - Top 10 Worst Impacted Nations in terms of COVID stats
- *Dashboard < Weekly/Daily Trends Prediction>*
  - Predictive Model to Show Daily/weekly/Monthly COVID trends
  - Video Feature to Showcase Predictive Analysis
- *Dashboard <General Customised Announcement>*
  - Latest Guidelines from WHO Regarding the COVID 19
  - Latest Safety measure issued by WHO
  - Latest Guidelines from State/Government regarding the COVID 19
- Dashboard < *Prediction Modelling Trends*
  - Predicted 15 days of COVID 19 data
  - Total Count of worldwide Affected COVID confirmed cases
  -

3.4 Target Audience

The Most important area to implement the project solution was to identify the Target Audience which can benefit most from the provided features. Our Project aim was to cover the affected population as well as affected government bodies which had to undergo a lot of extreme pressure and critical situations.

- *General Audience (World Wide)*
  - The Local people who are not able to get hold of Daily data
  - Country wide COVID Impact
  - Check Daily/Weekly Count of their country
  - Check WHO Guidelines
- *Government Bodies (Individual Country)*
  - Use the Dashboard for active measures on COVID 19
  - Customise the dashboard for preparation of State/Local announcement
- *Healthcare Sector*
  - Use Dashboard for readiness of medical facility (as per COVID count)
  - For future preparation of Medicinal Consumption, arrangement of Hospital bed and medicines
  - Updating latest guidelines by WHO which can help to revise the safety and precautionary guidelines

4.Experiment

To achieve our aim for this project, we aspire to model the collected data and perform machine learning algorithms to predict the outcomes which can be further used to visualize and compare the growing trends of the virus. Data science is a process which involves multiple steps ranging from data collection to data prediction and comparisons. Using the Data Science project, we aim to collect and model the data and infer the outcomes for the benefit of the project conclusions. The essential steps to achieve the same include Data collection, Data Preparation, Exploration, Modelling and Evaluation.

❖ Data Collection:

To provide prominent data analysis and create an impact, a good source of data is required. In this project as we aim to provide a prediction of corona virus cases on daily basis, we needed a data source that will provide daily cases in the world as well as in every possible country.

After exploration the web, we came across various data sources, we decided to go with the John Hopkin's university data set. The data is available for free use on the Git Hub website listed below. Based on this data source, we could explore and model the data. Hence, we collaboratively decided to go ahead with this data source. More details of the data source can be found here.

❖ Data Preparation:

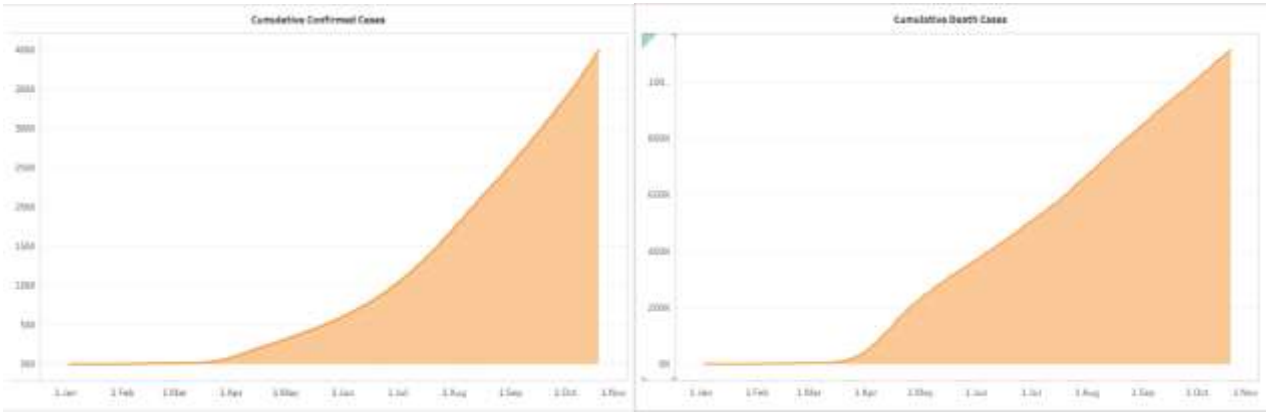
Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. This involves checking the format of data, re-formatting and checking the cardinality constraints, removal of irregular cardinalities is a step performed in this step. It is an essential step as to put data in context in order to turn it into insights and eliminate bias resulting from poor data quality.

In the obtained data set, we observed that the data was collected over time. To observe and remove any irregular cardinalities, we use python language and the python libraries to check and remove if any. We observe that the data is indeed a time series format, hence we convert the data into appropriate time series format. This is particularly performed to ensure veracity in data.

❖ Data Exploration

Data Exploration is an approach to perform initial data analysis. This is an important step in the data science process as the initial exploration provides some insights of data which can help us understand relationship between features and characteristics of data.

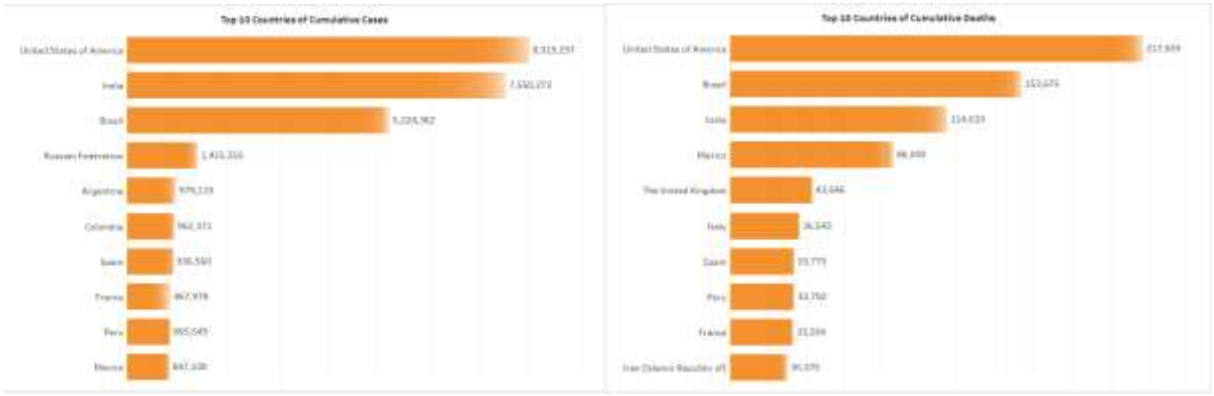
In this project to gather more insights about the collected data, we utilized tableau software to visualize.



Worldwide Confirmed and Death Cases

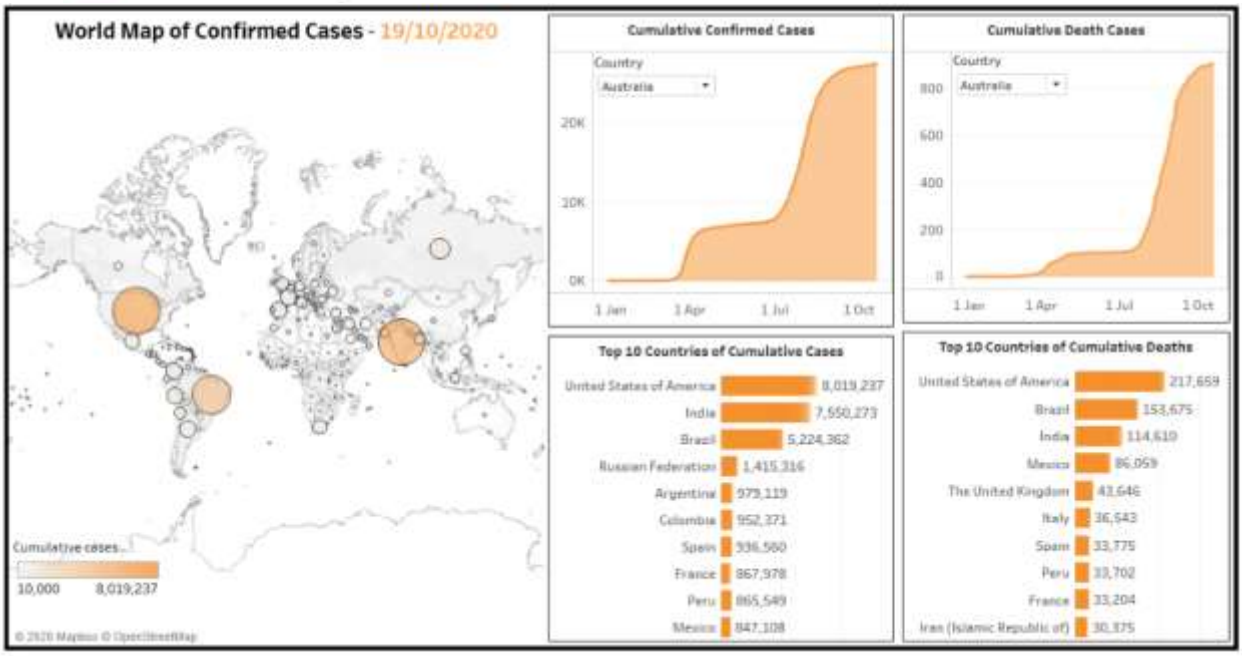
The above graph displays the confirmed and the death cases over time on a global level. It is evident from the above graph that an exponential increase in the confirmed and the death cases can be observed.

The number of cases for top 10 countries and the death cases as of 19th October 2020 can be observed below

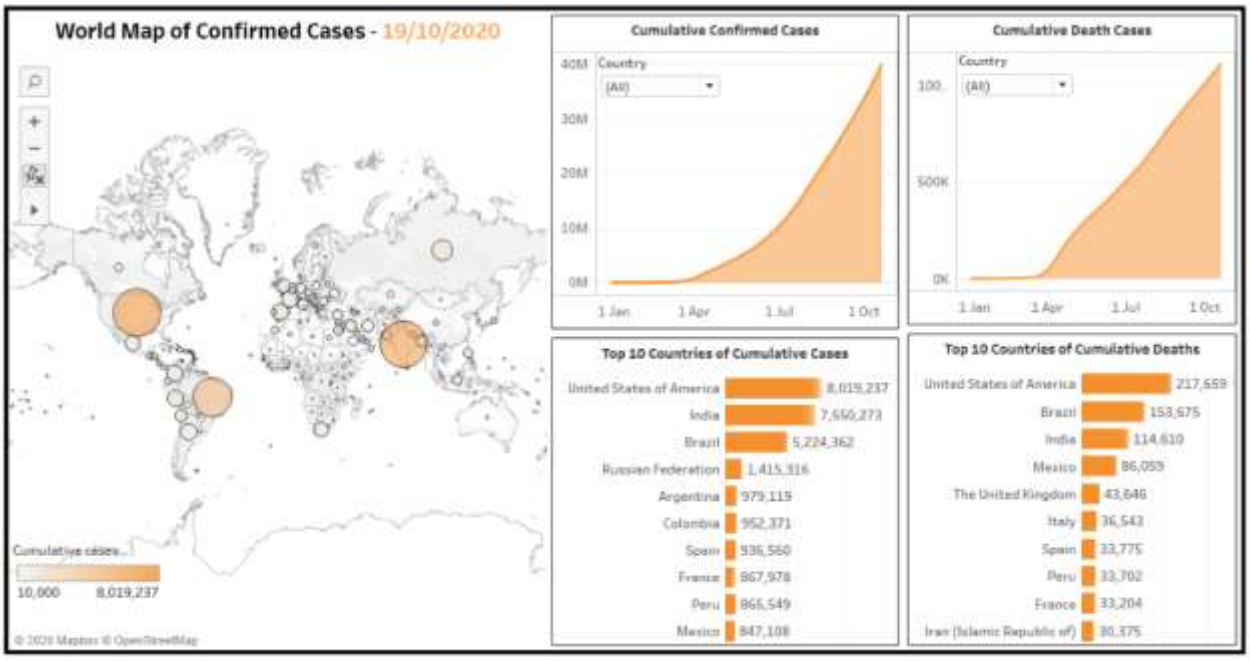


Number of cases and deaths for top 10 countries

Hence to extend our abilities and to infer more insights of the data, we used the software to prepare an application, a visual dashboard that displays all the information country wise and globally. It also helps us understand the increase with time as it produces a timely graph on global level. The following figures depicts the use of Tableau with a global and Australia's data. Similarly, data for each country can be displayed visually.



Number of cases and Death Rise in Australia



Number of cases and deaths rise worldwide

Based on the above visualizations it can be deduced that an exponential increase in number of cases can be worldwide, whereas for Australia we can observe bi-model graphs describing the first and the second wave of the virus. Using this information, we endeavour to prepare a machine learning model to predict the number of cases globally. Similar use can be made for country specific data.

4.1 Data Modelling

In this phase of data science process, we endeavour to predict the rise in cases daily for the next 15 days. These objectives can be achieved by pushing the cleaned data on a machine learning platform, devising and tuning a machine learning algorithm to obtain accurate and appropriate results.

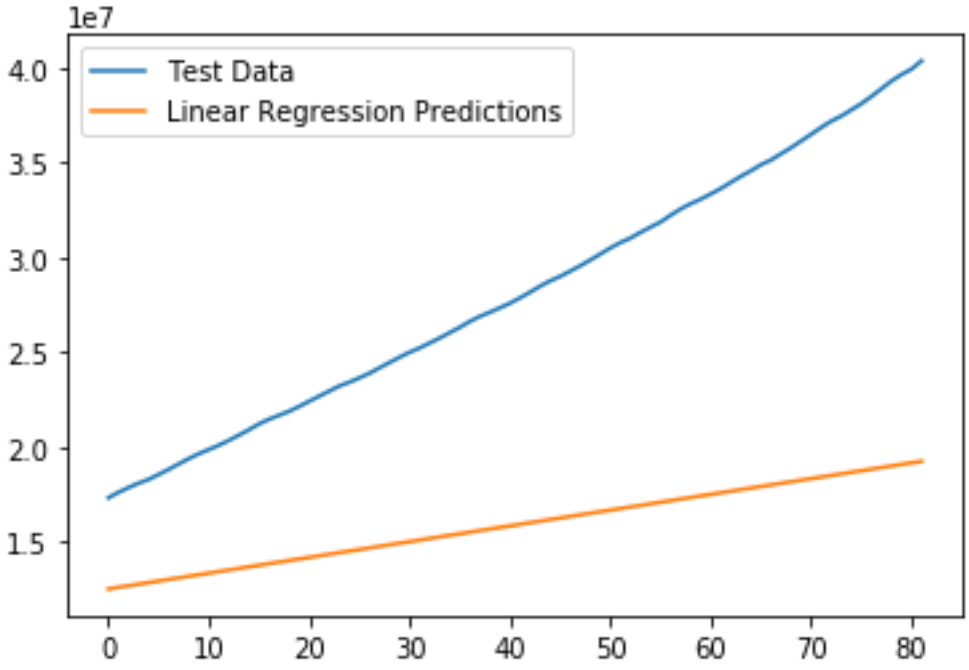
These results can be useful to gain insights and provide further analysis based on the obtained numbers of cases that are rising on a 15-day span. The similar observation can be done for a monthly or weekly basis and for individual countries and states or cities can be fed as the data input.

In the data exploration phase, we observe that the data collected is real numbers of cases diversified by the region. In the exploration phase, it is evident that the exponential increase in the number of daily cases can be observed. To use this information and build a model to perform the analysis on the collected data, we decide to go with regression algorithms for modelling.

Prior to modelling the data, we decide to train and test the model on the available data by splitting it into 70% training and 30% testing data. By doing so, we ensure that the model is consistently trained on the training data and is tested on the unseen testing data. This indeed helps us understand the performance of the model by comparing the actual values and the obtained results.

4.2 Linear Regression

Linear Regression was our first approach towards modelling the data. The linear regression model follows a linear relationship between the dependent and the independent variables. The dependent variable is the scalar response variable, and the independent variables are the feature variables. While the feature variables are used for exploration and explanation of data, the response variable records the target outcome of the relationship. To explore this relationship and to model our data, we use python libraries to develop a linear regression model.



Test Data and Linear Regression Predictions

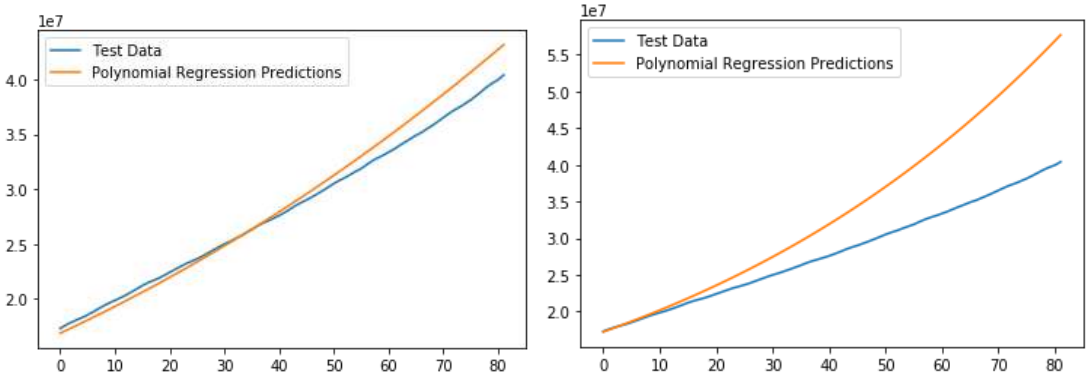
We modelled our data using the linear regression model on the world number of cases. The above figure demonstrates the results of the predictions and the actual values on the unseen data. The MAE (Mean Average Error) and MSE (Mean Squared Errors) are used as a measure to have an insight of the data model. It can be observed that the difference between the Test data and the Linear regression predictions is huge and hence we can deduct from this experiment that we require a better model for representation.

4.3 Polynomial Regression

Since we observe the inefficiency of linear regression model, we obtain a solution to go for polynomial regression. Polynomial regression is a form of regression analysis which models the relationship between the dependent and independent variables modelled as an n^{th} degree of polynomial. Since the exponential increase in cases are observed and the data represents the same, we go for a polynomial regression model. Since we have established that the data is a time series representation of the number of daily cases, we assume the high correlation between the features, also the relationship is no longer represented accurately by the linear regression model. Hence, we develop a polynomial regression modelling and try to fit the data in a polynomial equation.

In this project, we used the degree of polynomial as the hyper-parameter while tuning our model. Using the provided information, we built a polynomial regression model on the 70% training data and then tested the model on the remaining 30% unseen data. We obtained results using 3- and 4-degree polynomials which was satisfactory outcome.

In the below figure it is evident that the results of the above analysis provide a satisfactory outcome for the prediction model. Moreover, to test our model we checked the MSE (Mean Squared Error) and MAE (Mean Average Error) and reduced their error rate to obtain more accurate predictions. These predictions were obtained on 3- and 4-degree polynomial, hence we decided to take an average of these values and provide solution with the most accurate values.



Polynomial Regression Predictions using 3- and 4-degree polynomials

The following figure illustrates the polynomial predictions on the test data following an exponential path. The graph also shows the difference between the predictions of the linear regression and the polynomial regression predictions. It is evident from the visualization that the predictions of the model on the unseen data using a polynomial regression following an average of degree 3 and degree 4 model produces an outcome having most accurate predicts.

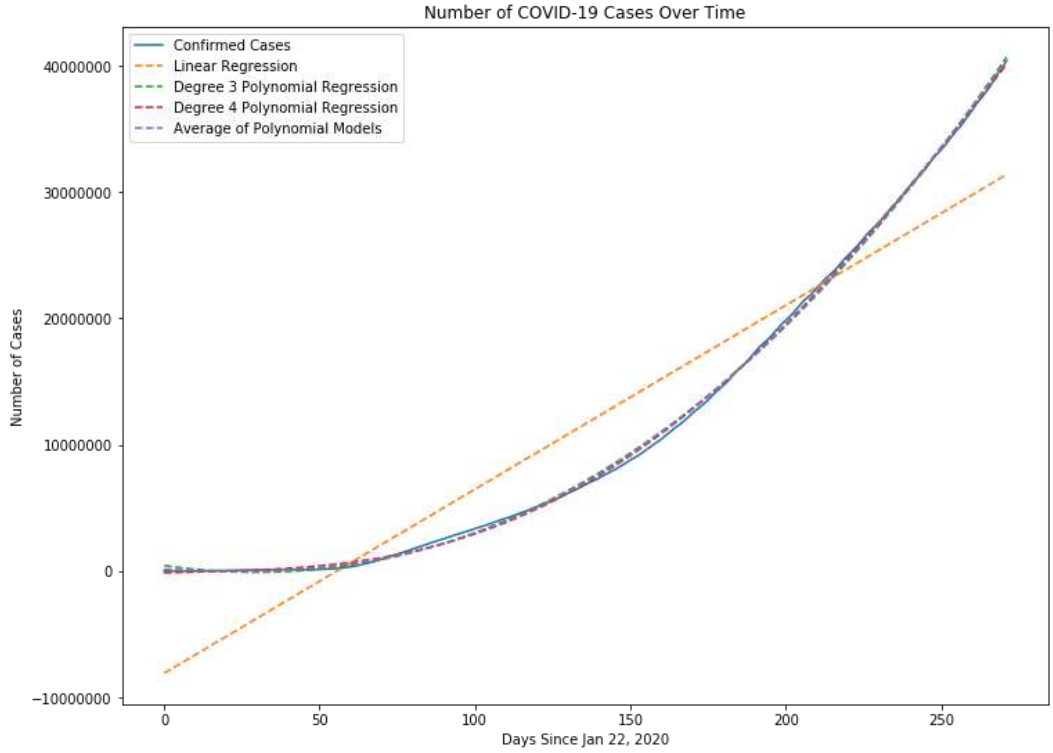


Figure showing the polynomial regression prediction and linear regression predictions

We can deduce this information from the above figure that the cases rise exponentially with time and the predictions on the unseen data that is the test data are accurate. Hence, we endeavor to use this model to provide a 15-day prediction of number of cases and the rise in cases.

5.Results

With the tuned machine learning model using polynomial regression, we predicted a 15-day outcome of the daily cases on a global level. These predictions can be observed in the following table. In the table the date specific values are displayed. Hence these values can be used as a conclusion of our experiment.

	Date	Number of worldwide cases predicted
0	10/20/2020	35750616.0
1	10/21/2020	36074744.0
2	10/22/2020	36400056.0
3	10/23/2020	36726544.0
4	10/24/2020	37054198.0
5	10/25/2020	37383011.0
6	10/26/2020	37712972.0
7	10/27/2020	38044073.0
8	10/28/2020	38376304.0
9	10/29/2020	38709655.0
10	10/30/2020	39044118.0
11	10/31/2020	39379682.0
12	11/01/2020	39716338.0
13	11/02/2020	40054075.0
14	11/03/2020	40392885.0

15-day Predictions using Machine Learning Model

6.Limitations

In the above section we experimented with some Machine Learning techniques on the data to check whether it is possible to predict the number of cumulative cases of COVID-19 confirmed patients for next 15 days. We used some basic ML algorithms to perform this such as Linear and Polynomial Regression models. Though we are successful in predicting the COVID confirmed cases for the next 15 days there are many limitations to this.

1. While performing experiments in the above section, we considered that the rate of increase in the confirmed cases will remain as it is without any sudden changes.
2. We do not consider any external factors like new government policies or new results in the COVID vaccine invention which are going to affect the rate of increase in confirmed cases all over the world.
3. For the demonstration purpose we considered the world data for our experiments. The developed model may or may not work with every country's data because the trends in increase of confirmed cases are different in every country and our proposed model may not fit all these trends.
4. The predicted result is just an experiment whether we can perform predictive analysis on the COVID data, and the results generated from it may not be completely true or valid.
5. Use of Linear and Polynomial Regression is the easiest way to experiment predictive analysis on the above data. Though we can experiment this with the use of more accurate and related forecasting techniques like 'Time Series Forecasting' methods like ARIMA and Neural Networks based deep learning techniques.
6. Each country has its own set of regulations and norms related to the COVID19 policy and our model does not take that into consideration as we are considering the whole world confirmed cases to create our model.

7. As the data is sourced from COVID-19 Data Repository by the Centre for Systems Science and Engineering (CSSE) at Johns Hopkins University, we cannot guarantee the truthfulness or integrity of the data used in this project.
8. The above experiment is just made for the academic purpose and it may or may not be used in real time applications with or without any additional changes in it.

7.0 Conclusion and Future Scope

As mentioned earlier, the problem and the concern raised by the COVID 19 has really taken the world to critically think about the future. There are many sources and organizations which are working so hard to alleviate this issue in the coming future. The good news is the data related to the COVID19 is getting published on real time basis by many sources. And with the help of advanced data analytics processes and the power of the data prediction we can obtain some interesting results from a data which are useful in solving some important issue mentioned in the problem statement.

Exploring different patterns from the data can help us deciding our future policies as a private or public sector organization. It can help scientists to accelerate the process of development of the potential vaccines. It can help hospitals and governments to arrange additional medical services and develop new policies.

We as data science students tried to explore this critical area through some data science practices and exploration and analysis techniques. We tried some simple Machine Learning techniques to check whether we can predict the future data on the basis of the current trend. We observed that in the future, with the help of other more accurate ML techniques we can solve this problem for each region, each country and even for the state level. Use of data visualization and exploration plays an important role in the data science process and in the future, we will try to focus more on the exploring data deeper and deeper to find out some interesting and useful trends from it which can then used for solving many problems with further data science practices. Although, it is difficult, it is not impossible to eradicate the effect of the COVID-19 on the world and the data science/analysis is playing an important role in this fight.

8.0 Project Management Description

Covid-19 project was completed as a result of collaborative efforts of 4 team members. This project provides a platform which describes an extensive use of Data science process and provides an overview of the number of cases and predictions parallel to the daily increase in cases globally. To achieve our aims, we worked together dividing the tasks based on individual expertise and helping each other collaboratively to endeavor representative results. The members of the project group include, Gaurav, Pranamya, Tanmay & Shubhankar.

The complete project was divided mainly into 4 phases namely, initiation, planning, execution & closure. With these division of phases, we could achieve a small feedback within the team at the end of each phase which would highlight our strengths and weaknesses. To co-ordinate efficiently and have a track of the project development measures, we used numerous open source tools such as trello, Microsoft teams, GitHub repositories, Prezi and Google Docs. The significance of each tool as we advance with the project is discussed below.

8.1 Initiation

In this phase of project development, we collaboratively discussed and finalized to advance with COVID – 19 as our project. In this phase we discussed about developing a business problem on which we can advance based on the data science processes and provide an efficient solution. Microsoft Teams was extensively used to set up meetings and discuss about the findings.

In this phase we further discussed data collection techniques and sources of data. While Pranamya and Shubhankar explored the available open source visualizations, Gaurav and Tanmay explored various data distribution platforms like UCI Machine Learning Laboratory and Kaggle. Interesting development of the project was posted on Trello, hence, we were able to keep track of the individual achievements and reduce executing redundant tasks.

As we progressed through the initiation phase, with continued collaboration, in discussion calls we came to a solution of having real time data of everyday count of corona virus cases. While Gaurav and Tanmay produced some meaningful data sets and a survey of such data sets we finalized John Hopkin University's real time cases data. Shubhankar and Pranamya studied the visualization published by bing.com which is a daily corona virus tracker.

Based on the study and the data collected in this phase we finalized our aim to develop a strategy to implement prediction algorithms and develop an interactive platform to produce representative visualizations for coronavirus daily affect count on a global scale.

8.2 Planning

In this phase of project development, we extensively used Trello to disintegrate problem statement into smaller chunks of tasks and assign each task to teams as so to monitor and provide timely feedback.

In planning phase of project development, we collected a series of steps to perform to achieve our aims. In this phase we planned our progress of the project with respect to the submission duration. Phases of Data science process were divided according to the project requirement and individual tasks were assigned based on the expertise and domain knowledge.

Trello software was used to calculate and monitor progress and provide individual feedbacks.

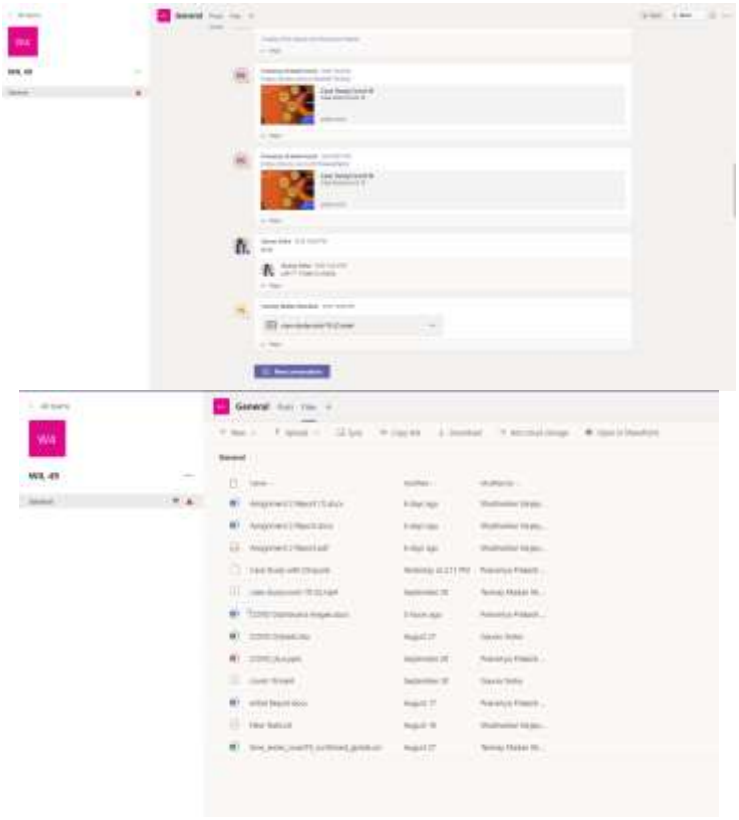


Figure demonstrating the use of Teams Application

8.3 Execution

Once the project was planned and the individual tasks were divided, we were good to execute the experiment and perform analysis. Pranamya and Shubhankar studied various Machine Learning algorithms and techniques which can be efficiently used to explore and train the data. It was found by Pranamya that the daily timeline data was indeed a time series problem and could be solved as a classic regression problem. Association of the dependent and independent parameters was explored by Shubhankar, while Gaurav and Tanmay worked collaboratively in data extraction and cleaning.

To track our progress with the experiment we used GitHub as the platform. Using GitHub we could efficiently monitor the progress of data cleaning and exploration produced by Gaurav and Tanmay. Pranamya and Shubhankar were able to clone the repository and work individually on building regression models. The Machine Learning models were built in Python, using Jupyter notebooks (Anaconda 3). While Shubhankar worked and produced results using linear regression model, Pranamya found an interesting trait in the data and extended the experiment to use polynomial regression methods.

Once the model was efficient and trained on the training data, we worked collaboratively towards the model testing. Since we had an online data source, we were never deficient on availability of data. Hence, each member worked on the model testing with unseen data. Pranamya collected and mapped all the results to study the model performance. Based on the model performance, Shubhankar and Pranamya tuned the model and was pushed to GitHub repository. Gaurav and Tanmay tested the tuned model and with the approval of all the team members we produced the outcomes and concluded our findings.

8.4 Closure

In this phase of project development, we aimed to communicate our findings as a solution to the defined business problem. Gaurav and Tanmay worked together to find a platform to communicate using good presentations and reports.

Presentations were made on Prezi by Gaurav and Tanmay. These included various visual representations. These presentations were then recorded describing various project outcomes. We then used Google docs to divide and present report so that the outcomes of the project are well represented.

9.0 References

1. Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 2020; doi:10.1056/NEJMoa2001316.
2. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020; 395: 497–506.
3. Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art, 2020
4. Propagation Analysis and Prediction of the COVID-19, 2020.
5. Perez Perez, G. and Talebi Bezmin Abadi, A., 2020. *Ongoing Challenges Faced In The Global Control Of COVID-19 Pandemic*.