Machine Learning Techniques for Trend Prediction of Epidemic-like Disruptive Events and Their Application wrt Specific Environment

by

Saman Bashir

A thesis submitted to the faculty of Information Security Department, Military College of Signals, National University of Sciences and Technology, Rawalpindi in partial fulfilment of the requirements for the degree of MS in Information Security

July 2021

CERTIFICATE

Certified that final copy of MS Thesis written by **NS Saman Bashir**, (Registration No. **00000203537**), of **Information Security Department, Military College of Signals** has been vetted by undersigned, found complete in all respects as per NUST Statutes / Regulations / MS Policy, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS degree. It is further certified that necessary amendments as pointed out by GEC members and foreign/local evaluators of the scholar have also been incorporated in the said thesis.

Signature:

_____

Name of Supervisor **Asst Prof Dr. Malik Muhammad Zaki Murtaza**

Date:

_____

Signature (HOD):

_____

Date:

_____

Signature(Dean/Principal)

_____

Date:

_____

DECLARATION

I hereby declare that no portion of work presented in this thesis has been submitted in support of another award or qualification either at this institution or elsewhere.

DEDICATION

This is dedicated to my beloved parents for their unconditional love, endless support and encouragement.

ACKNOWLEDGEMENTS

I am thankful to Allah Almighty who has given me strength and firmness to accomplish this research. Without His consent, I would not have been able to complete this thesis. I am highly grateful to my supervisor, Dr. Malik Muhammad Zaki Murtaza and co-supervisor Asst Prof Dr Saddaf Rubab for their worthy supervision, continuous support and valuable help throughout this study. Also, I am highly thankful to my committee members Asst Prof Mian Muhammad Waseem Iqbal, and Asst Prof Dr Yawar Abbas for their valuable suggestions. Finally, I am grateful to my friends and family for their continuous support.

ABSTRACT

COVID-19 disease has almost been witnessed by all countries of the world in 21st Century and is now declared as pandemic by World Health Organization (WHO). It is still a growing challenge due to testing of vaccines to cure this disease and it can be predicted that this process may still take quite a lot of time in providing to masses. In Pakistan, first case of Corona virus was reported on 26 February 2020, afterwards COVID-19 infections spread at quite a high pace in the year 2020. In 2021, COVID-19 is still a growing challenge for Pakistan and strict measures needs to be taken by health sector of Pakistan to abate the risk of communication.

Machine learning techniques are used to help describe the size of an outbreak and the rate of spread of an infection in a population. Some Machine learning techniques were employed, around the world, to estimate the severity of the COVID-19 pandemic. Trend of this disease has shown varying behavior in almost all affected countries, therefore 1 method suitable for a specific country will not likely work for another country. It has been proved from deduced graphs that different degrees of polynomial regression model are working for different countries COVID-19 data for the same time period i.e. 13 months. Moreover, COVID-19 curve of a country has massive fluctuations, which depicts various feature sets are playing their role, leading to increase and decrease of cases. Therefore, there is a need to evaluate Government policies, Environmental and cultural feature sets playing their role in rise and fall of severity of disease. In Pakistan, it's been 1 year still the country is not out of pandemic dangers, so we feel it is a good time to study the trend of

feature sets which have affected the trend criticality to reflect on the predictions of pandemic spread.

In order to build SOPs and future measures there is a need to observe and analyze the Machine Learning (ML) techniques for COVID-19 trend prediction and compare them on basis of performance parameters. This work will not only help the country's on-going efforts towards controlling the coronavirus, but it has also proposed a framework which will work as the basis of any prediction systems for such disruptive events in our society.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background

In December 2019, a life threatening infection was identified in one of China's city Wuhan [1] and it was highlighted that this virus was originated from Bat [2]. It took almost 213 countries in its fold by 04 June 2020 infecting around 6.63 million people and 0.39 million deaths. It became hard to detect at first because this disease has very close symptoms to pneumonia. Due to massive deaths and high spread rate WHO declared this disease as pandemic in March 2020 [3-8]. It is a respiratory disease and has symptoms like cough, diarrhea, flu and fever [9]. The existence period of virus is around 12 hours and most common in elderly 60 years and above but now it is also becoming viral in children [10-11].

1.1.1 Epidemics Observed in Last Two Decades

Various epidemics similar to corona virus have been observed in last two decades, however their infection rate was quite lesser than COVID-19 but death rate was varying as shown in Table.1.

Epidemic/Pandemic	Start Year	End Year	Cases	Deaths	Death Rate
SARS	2003	2004	8096	774	9.56
EBOLA	2014	2016	28646	11323	39.53
MERS	2012	2017	2494	858	34.40
HINI	2009	2010	6724149	19654	0.29
COVID-19	2020	-	6630000	390000	5.9

Table 1: Summary of Virus Observed in Last 2 Decades

One of the disease similar to COVID-19 was Severe Acute Respiratory Syndrome (SARS) reported in 2003 which was spread a lot and cause many deaths. Middle East

Respiratory Syndrome (MERS) emerged in 2012 and stayed for around 4 years causing number of infections. Similarly, EBOLA and HINI outbreak were observed in 2014 and 2009 respectively which also caused quite a havoc [12-14]. However, it has been observed that COVID-19 spread rate is also quite high as by 16 March 2021 confirmed cases reached 15484 million and deaths by 342 million globally.

1.2 Global Situation of COVID-19

COVID-19 virus is continuously growing and it has affected economy of even developed countries greatly. The world has observed second corona wave in which confirmed cases have reached 15484 Million and deaths reached 342 Million. Confirmed cases and deaths have risen since identification as shown in Fig 1 and now third wave is ongoing. However, it has been observed that this disease has affected every country differently and cases are varying. Few countries are severely affected and some have quite stable situation. Due to existence of many variants of this disease, the cure to this disease is taking a lot of time.

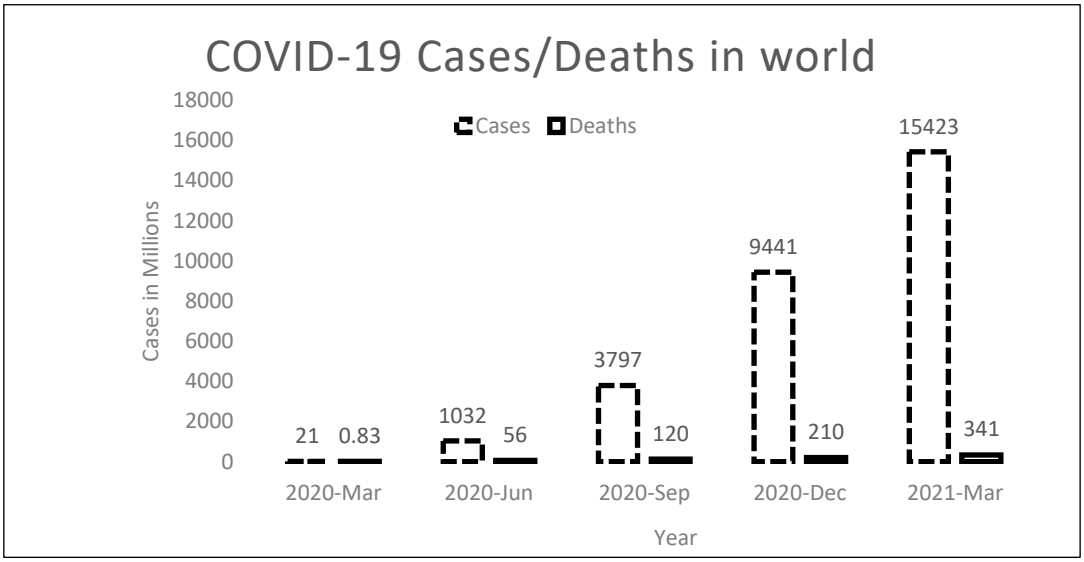


Fig 1: COVID-19 Around the World

1.2.1 Severely Affected COVID-19 Countries

Countries such as United States, Italy, Spain and UK are severely affected by this disease and their confirmed cases and deaths rose magnanimously as compared to other countries. The cases of each country (In Millions) are shown in Fig 2 [15]. Even in severely affected countries it has been observed that curve is varying and none of the two countries are following the same trend. This employs that future prediction of cases in all countries will be different and different ML methodologies to be used for every country to analyze the disease curve.

1.2.2 Less Severely Affected COVID-19 Countries

There are few countries which were not critically affected by COVID-19 or affected for a limited time period; on the basis of cases and death rates recorded as compared to other countries. Some of the countries include Australia, New Zealand, Bhutan and Fiji. At present, in these countries confirmed cases and death rate is almost close to zero as shown in Fig 3 [15] which depicts these countries have different affecting feature sets.

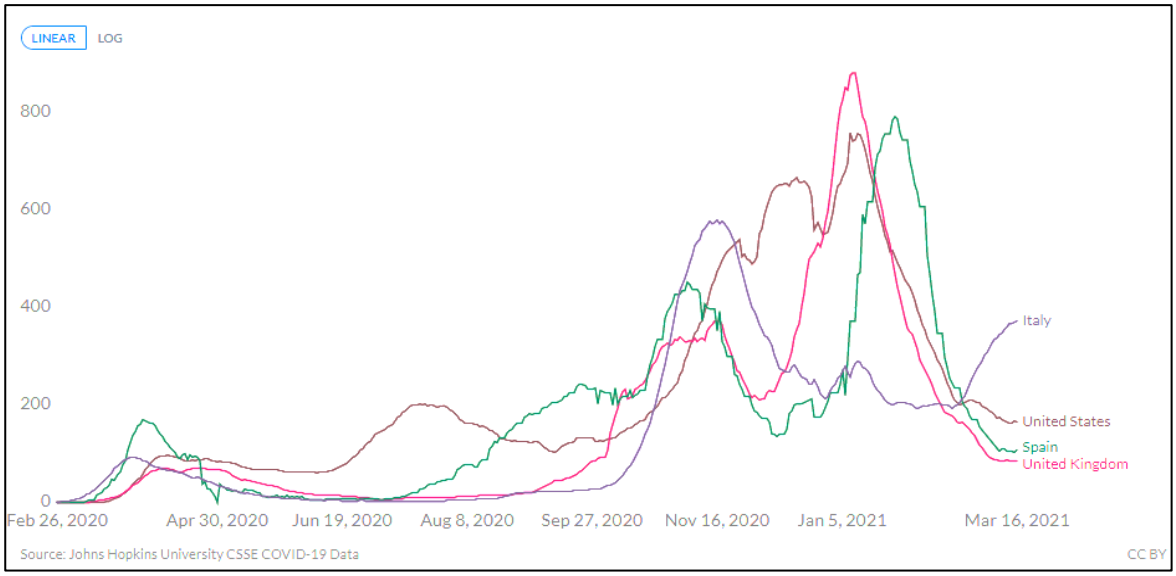


Figure 2: COVID-19 in Severely Affected Countries of the World

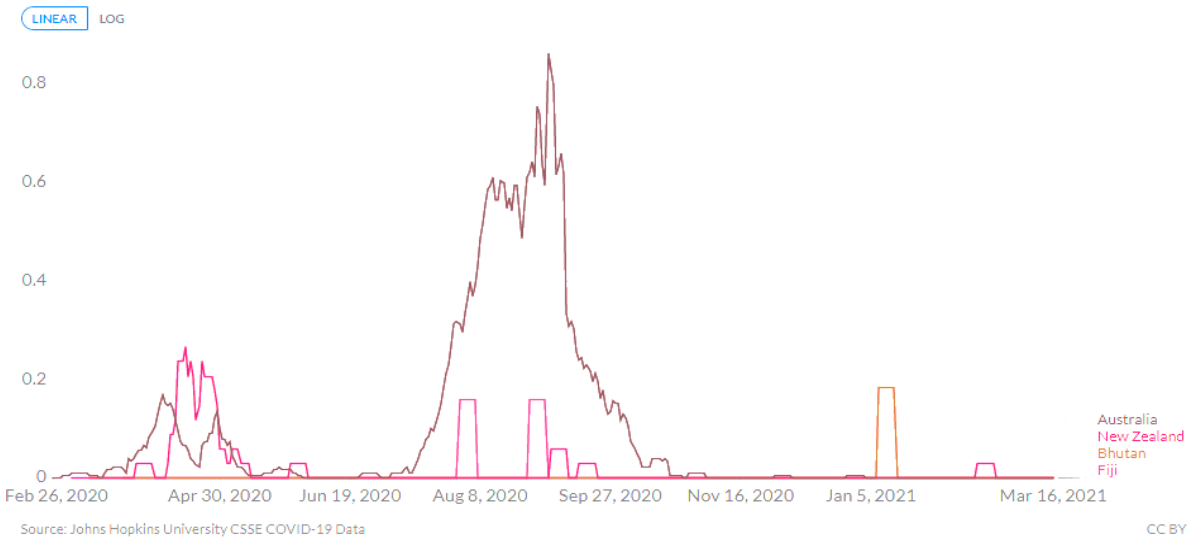


Figure 3: COVID-19 in Less Affected Countries of the World

1.2.3 Pakistan

On 26 Feb 2020 COVID-19 came to Pakistan which then spread at quite a high pace uncontrollably [16]. It has affected almost all sectors of the country either its education sector, economic, banking etc. Pakistan COVID-19 curve is highly varying as shown in Fig 4. Sometimes the cases were reduced a lot which depicted that disease has been controlled, however, after first wave, second wave was observed which lead to ongoing third wave, thus resulting in to critical situation in Pakistan including high mortality rate [17].

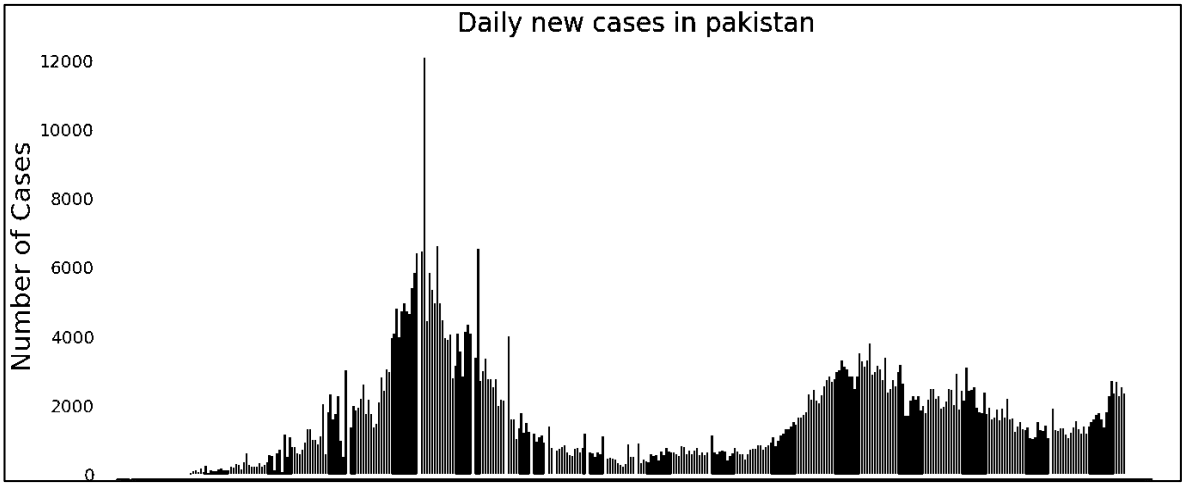


Figure 4: COVID-19 Daily Cases in Pakistan

1.3 Motivation:

From Jan 2020 onwards corona virus has been a challenge for the whole world and specifically for Pakistan. COVID-19 has greatly affected every sector of Pakistan including unemployment, health system and education. The novel corona virus is new and researchers all around the world are analyzing the datasets to design a model or suitable method for future predictions of this disease. Since, trend of corona virus is varying rapidly it has arisen the need of trend prediction and analyze the features which are making data varying in almost all countries and even within country in different cities.

It is necessary to do future prediction of COVID-19 to understand the growth of cases, mortality rate and features which have increased or decreased cases at a massive level in Pakistan. This research will provide insights to predict the dynamics of COVID-19 in near future and suitable ML based technique related to Pakistan environment. The prediction will help government and health sector to take measures accordingly to control the increasingly challenging situation or lessen the challenges faced by government and society due to this disease.

1.4 Problem Statement

COVID-19 virus has spread in almost whole world in the year 2020. In February 2020 it started spreading in Pakistan, since then it has risen continuously. After first wave, second wave was witnessed by country followed by an ongoing third wave which has increased the challenges for Pakistan. Therefore, strict measures needs to be taken by health sector of Pakistan to abate the risk of communication and control its spread. ML techniques and models have been used all around the world to describe spread of this disease and predict the spread and mortality cases ahead of time. Researchers

have applied different techniques majorly on cumulative data instead of daily cases which is not an optimized solution for varied data set.

It has been observed that this disease has affected every country in a distinct manner in terms of positive cases and death rate, therefore, model derived for one country may not be suitable for another country. COVID-19 curve has a lot of variations due to increase and decrease of cases, so research done for specific time period will not be affective for large time period with. Also, all countries have different government policies, environment, cultural features which have acted upon disease curve, which is also one of the crucial reason of variations in trend.

Since, the disease spread in Pakistan has not still been reduced in 2021, hence it is required to develop a detailed understanding on how to best model the country's trend including the problem areas discussed in the previous section. Therefore, an efficient and proactive ML approach is required keeping in view time and geographical area bounds.

1.5 Project Description

1.5.1 Objective

Objectives of the thesis are as follows:

- Selection of ML technique to analyze and predict COVID-19 trend using Key Variables i.e. confirmed cases, deaths and recoveries
- Comparison of selected technique variants on the basis of performance metrics
- Examination of techniques on dataset of other severely affected COVID country (e.g. Italy) and comparatively less affected (e.g. Australia) country
- Comparison of results of other selected countries with Pakistan to examine suitability of ML technique w.r.t geographical area

- Identification of environmental, cultural and government policies feature-sets which can explain varying disease curve of Pakistan
- Propose a framework to automate the selection of ML technique along with inclusion of feature sets and its impact on prediction of data

1.5.2 Approach

Our approach has been shown in figure 5.

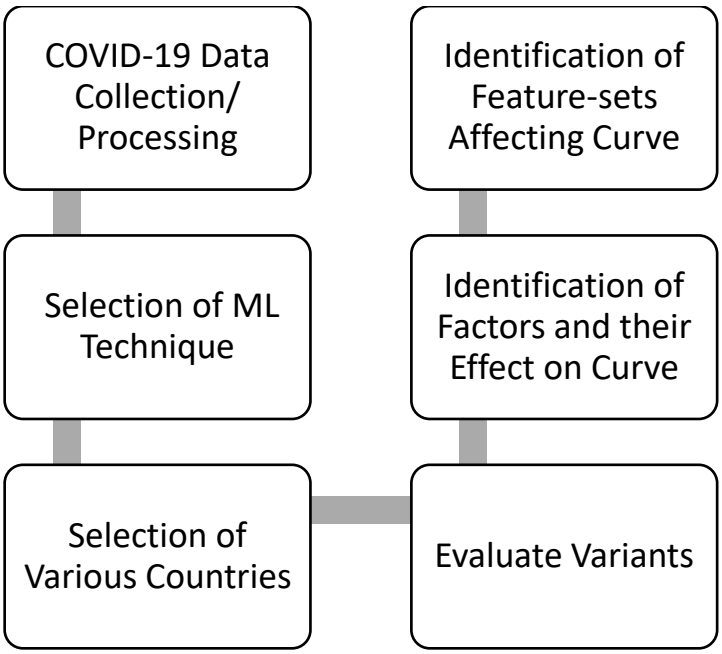


Figure 5: Approach of the Proposed Research

1.5.3 Areas of Applications / Advantages

This research will be useful in

- Health sector, behavioral sciences departments, emergency clinics and organizations working on COVID-19
- Help in detecting ML technique best suited for a country to foresee future dynamics

1.6 Outline

This thesis has been divided into 6 chapters:

Chapter 1: This chapter introduce our thesis. It explains the background followed by proposed approach. Also, it explains objectives and application areas of the thesis.

Chapter 2: It includes related work done in the same dimensions and ML techniques used by researchers worldwide and Problem area.

Chapter 3: This chapter includes Selected ML technique, proposed methodology, hardware and software setup to achieve above mentioned objectives.

Chapter 4: It deals with deduction of results and evaluation of major feature sets.

Chapter 5: It includes analysis of results and a proposed framework.

Chapter 6: This chapter concludes the thesis along with future directions.

Chapter 2: Related Work and ML Techniques

As it has been discussed earlier that COVID-19 behavior has been dissimilar in various environments. Therefore, in order to study its behavior many researches have been done on this disease all around the world to predict it for the future.

2.1 Related work

Raji and Deeba in 2020 [18] performed a study to analyze the effects of COVID-19 in various cities of India. Cumulative dataset was taken from kaggle website from 2 Mar 2020 to 28 July 2020 and modelled using linear and polynomial regression techniques. The dataset of different cities of India was analyzed and variables were predicted for future to give measures for handling of disease in a better manner. They focused majorly on death cases of 7 cities and concluded whether cases will increase or decrease in specified area.

In 2020 [19] Rabia, Sadia, Asif, Amir conducted a study on Pakistan COVID-19 data taken from World Health Organization (WHO) 26 Feb 2020 to 3 Apr 2020. Analysis was done by collecting daily data of tests performed, confirmed cases, deaths, recoveries and travelers screened using Statistical Package for the Social Sciences (SPSS) version 23. Also, future forecasting of confirmed cases in Pakistan has been done using Simple Moving Average technique. Finally, they concluded that cases will rise in future therefore SOPs needs to be formed to overcome spread of corona virus and stabilize the economy.

In 2020 [20] Spyros and Fotios performed the study on world cumulative data taken from John Hopkins website from 22 Jan 2020 to 11 Mar 2020. Confirmed, recovered and death cases were selected as variables for research and forecast has been done using Simple time series technique. Overall time data is divided in 5 rounds to analyse

rise and fall in COVID-19 cases with time. They concluded that COVID-19 will not only be confined in China rather it will spread and increase globally.

In 2020 [21] a research was done on USA cases by Rubaiyat, Subrato, Prajoy, Priya since, USA became worst hit country after China. They modelled cumulative worldwide cases using polynomial regression technique with degree 2 on data from December 2020 to 4 June 2020. Also, various classification techniques have been applied on data set taken from Hospital Israelita Albert Einstein of Brazil for diagnosis of disease.

Xiongwei, Hager, Eman, Radhya, Abdelmgeid [22] introduced a real time system on sentiment prediction of people on COVID-19 on Twitter. In their technique both offline and online modelling has been done on extracted data for the period 23 Jan 2020 to 1 Jun 2020. Test have been performed using decision tree, logistic regression, k-Nearest Neighbor (K-nn), random forest, and support vector machine learning techniques and best results were achieved using RF model.

Samia, Nyla and Zafar [23] designed a research on trend of COVID-19 in Eastern Mediterranean Region (EMR) with Particular focus on Pakistan. Data was collected from Pakistan National Institute of Health from 14 Feb 2020 to 26 Apr 2020. They studied trend of disease over time in 22 countries with particular focus on rise and fall of cases and mortality rate. By using SIR epidemiological model they concluded that number of cases will be half a million by June 2020, therefore development and implementation of SOPs must be ensured to tackle the seriousness of matter.

In 2020 [24] a study was presented by Aniele, Federica, Elisa, Giuseppe and Eloisa on analysis of COVID-19 in Italy and Italian Regions. Data was taken from daily reports from 20 Feb 2020 up to 81 days. Regression model, Gompertz growth equation was

used on cumulative data to analyze the confirmed cases. They finalized that testing of people and lock down in an effective solution in reducing further spread. They also studied some factors on a wider scale which affects COVID-19 cases. On local level they considered GDP, distance and interaction and at interventional level intercept, swabs and cases at lockdown were studied. Also, combination of both variables at local level and also at interventional level were studied.

Dominic, Mubarak and Manohar discussed the spread of Corona virus in Australia as it is one of the less affected country in 2020 [25]. Australia was able to overcome spread of disease due to strong implementation of strategic measures i.e. physical distancing, higher testing of individuals and geographical isolation. As first case was detected on 2 Mar 2020 and strict lock down was imposed on 13 Mar 2020. Similarly, working hours were also defined accordingly afterwards. Australia proved that since there was no vaccination to this disease they minimize the disease by their strict reliance on early management and policies.

A study was conducted in 2020[26] to analyze and forecast the cases of United States, Spain and Italy using hybrid polynomial Bayesian ridge regression model by Mohd Saqib. Cumulative dataset was taken from John Hopkins website till 11 May 2020 since identification. They proved that degree 6 model is best suited for Italy cases and similarly US cases are best modelled by degree 6. However, Spain cases are modelled using degree 7. Also, they presented that in future important factors affecting COVID-19 criticalities can be studied using similar technique.

Debanjan and Monisha performed a study for prediction of COVID-19 cases in India using support vector regression (SVR) in 2020 [27]. Data is collected from 1 Mar 2020 to 30 Apr 2020 and on the basis of their study they concluded that three to four months

will be taken to overcome the disease. Their model analyzed confirmed cases, recovered cases and death cases along with prediction of daily new cases of all three variables stated earlier.

2.1.1 Summary of Related Work

The selected related work in the domain of COVID-19 is summarized along with important factors in Table 2.

Paper	Area	Technique	Data Pd/ Type	Factors / *Pred Time
[18]	States of India	Linear & Polynomial Regression	2 Mar to 28 Jul 2020. Cumulative	Death Cases *6 Days
[19]	Pakistan	Simple Moving Average	26 Feb to 3 Apr 2020. Cumulative	Test performed, Confirmed, Deaths, Recovered, Travelers screened *1 Week
[20]	Global	Time series	22 Jan 2020 to 11 Mar 2020. Cumulative	Confirmed, Recovered, Deaths Cases *10 Days
[21]	US	Poly Regression; 2 Degree	20 Dec to 4 Jun 2020. Cumulative	Confirmed cases *1 Week
[22]	World wide	Decision tree, Logistic regression, k-NN, Random forest, and SVR	23 Jan to 1 Jun 2020. Tweets	Sentiment prediction of People on COVID-19
[23]	EMR, Pakistan	SIR Model	14 Feb to 26 Apr 2020. Cumulative	Confirmed, Death Cases *1.5 Months
[24]	Italy	Regression Model	20 Feb to 10 May 2020. Cumulative	Confirm cases, GDP, Distance, Interaction, Swabs and Lockdown *12 Days
[26]	US, Spain, Italy	Poly Bayesian Ridge Regression	Till 11 May 2020. Cumulative	Confirmed Cases *6 Days
[27]	India	SVR	1 Mar 2020 to 30 Apr 2020. Daily Cases	Confirmed, Recovered, Deaths Cases *1 Month

Table 2: Summary of Related Work

2.1.2 Analysis of Researches Conducted on COVID-19 and Identification of Problem Areas

Literature review that has been discussed in above section includes various techniques applied on different parts of the country in different time domains using distinct factors and variables. Form Table 1 it can be seen that usually researches have been conducted on cumulative data for the time period taken instead of daily data except Debanjan and Monisha's research. Although, analysis of daily cases instead of cumulative cases can give more in-depth view of increase and decline of COVID-19 trend at different times, which has further been covered by this research.

In most of the researches, variables mainly confirmed cases and deaths have been considered, however, recovered cases is also an important indicator to be studied to get an idea of recovery of disease in a specific country.

Feature sets such as environmental, cultural and government policies are significant features which have greatly affected COVID-19 trend, which are not found in any of the researches as per our knowledge except few variables studied in research paper [27]. However, analysis of these features can give more accurate prediction for future. These feature-sets have been studied in detail wrt Pakistan environment.

From related work it has been noted that [18] [21] [22] [24] & [26] have used regression techniques to analyze COVID-19 cases and perform prediction for different countries around the world for distinct time periods. Also, polynomial regression technique has degrees which makes this technique capable of modelling varying datasets. Since, our study also considers datasets of Pakistan, severely affected country and comparatively less affected country, therefore, this technique has been considered to

model the datasets as it has been widely used to analyze COVID-19 cases around the world.

Also, ML techniques used in different countries are giving varying results which are varying w.r.t selected time period as many features are playing their role in each countries pandemic trend accordingly. These issues which have been retrieved form related work provide us with an opportunity to perform a comprehensive study in this regard including detailed scenarios which will be discussed in upcoming chapters.

2.2 Concept of ML

ML is a branch of Artificial Intelligence (AI) focused on building applications that learn from data and improve their accuracy over time without being programmed to do so [28]. There are four basic steps for Machine learning as shown in Figure 6.

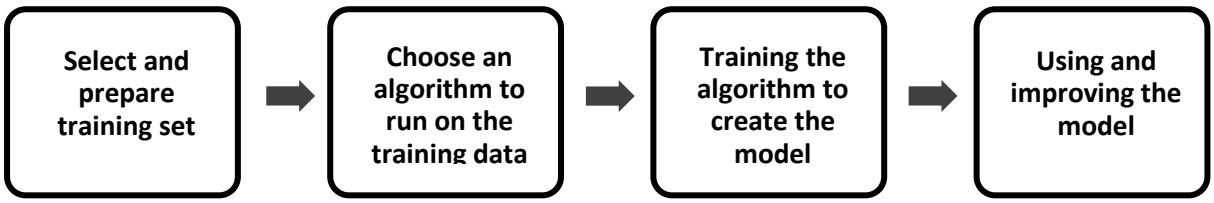


Figure 6: Approach of ML Techniques

Recent studies have proved that ML is a promising technology which expedites and help various organizations to analyze the trend and make scalable and reliable prediction and even outperforms human intelligence too [29]. According to American physiological society, the historic data of COVID-19 cases can be consolidated and analyzed by advanced ML algorithms to get a better hold on understanding of spread of viral disease, effective approaches for trend analysis and identification of associated factors [30]. Also, with the recent advancement in the field of ML and its computing speed and infrastructure, its horizon has been expanded in to the fields which were previously thought to be only human expertise [31].

2.2.1 Common Types of ML Algorithms

There are 6 broad categories of ML algorithms where first three are for labeled data [32] and last three are for unlabeled data [33].

Regression Algorithm [34]

Linear, polynomial and logistic, and Support Vector Machines (SVM) are termed as regression algorithms which are used to map relations in a dataset. Linear regression and polynomial regression is use for prediction of dependent variable value on the basis of independent variable. Logistic regression is used in cases where the nature of dependent variable is binary. Role of SVM comes in to play when classification of dependent variables is difficult.

Decision Tree Algorithm [35]

These algorithms give recommendations on basis of set of decision rules. It creates a training model which in result gives prediction of value of chosen variables deduced from training data. One of major subject algorithm is random forest. For prediction a root is considered which is start point, its value is compared with records attribute on the basis of which algorithm jumps to the next node.

Instance Based Algorithm [36]

These algorithms works on estimation that likelihood of a certain data point is closer to one group or with another on the basis of its proximity with other data points. It makes supposition from training values. It is also termed as memory bases learning. The complexity of this method is dependent on training dataset. However, one of its drawback is large data storage requirement and each query require starting the identification of local model from scratch. Its example is K-nn.

Clustering Algorithms [37]

In this algorithm there are certain groups and then groups of similar records are identified and labelled according to group to which they belong. The characteristics of groups is unknown beforehand. Example of clustering algorithms are K-means and TwoStep clustering.

Association Algorithm [38]

It works in if-then relationships after finding patterns and relationships in variables present in the data set. These relationships are called association rules. Matrices are used for finding the relationships between variables where support matrix shows frequency of variable in a dataset, Confidence matrix gives how many times if-then statements are found true and lift is use to compare real value with the expected value.

Neural Networks [39]

This algorithm defines and work in a layered network. Input layers takes data, hidden layer provides platform for performing calculations on input data and final layer is output layer which assigns probability to the conclusions drawn in hidden layer. If a network has multiple hidden layers they will give more refined results.

2.2.2 Polynomial Regression [40]

Polynomial regression is a form of linear regression which is used when data is highly nonlinear such as COVID-19 data. It is usually applicable when prediction can result into infinite possible answers. The independent and dependent variables have nonlinear relations and they can be fit using any suitable n degree of equation.

Therefore, it is believed that nonlinear data can be adjusted and can be fit in using any degree of “n” as shown in Fig 7. As it has been discussed earlier that COVID-19 data is highly nonlinear therefore this technique has been selected. In this study each selected country’s data has been modelled using different degrees of polynomial equation in which they fit and gives considerable performance parameters.

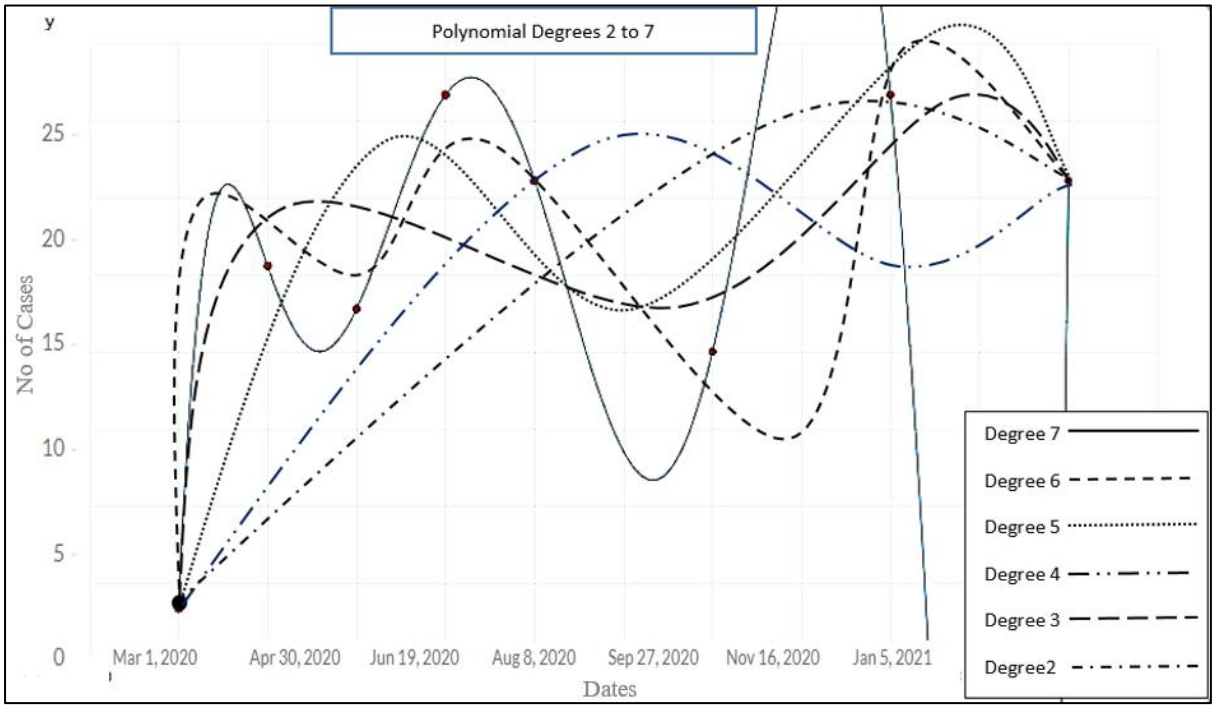


Figure 7: Trend Lines of Polynomial Regression Variants

Chapter 3: Our Methodology

In this chapter we will present our methodology to analyze COVID-19 trend and feature sets associated with it specifically for Pakistan environment and generally for few other selected countries.

3.1 Hardware/Software Infrastructure

All the tests have been performed on laptop using following:

Processor	Intel Core i5
System type	64-bit operating system
RAM	4 GB
Free Disk space	5 GB
Windows	10
Python	3.9.0
Pip	19.0
Software	Jupyter Notebook

Table 3: Hardware/ Software Infrastructure

Pip [41] is a packet manager for packages of python. Since, we are going to use python on jupyter notebook therefore we require pip. When Python setup is installed in a laptop pip get installed automatically however, it needs to be updated [42]. Simple built-in Command Line Interface (CLI) of windows are used for this purpose. After setup and installation, jupyter notebook will be opened in web browser, as shown in fig 8, which was set while updating Pip. Python notebook can be opened using new tab and code block will be opened.

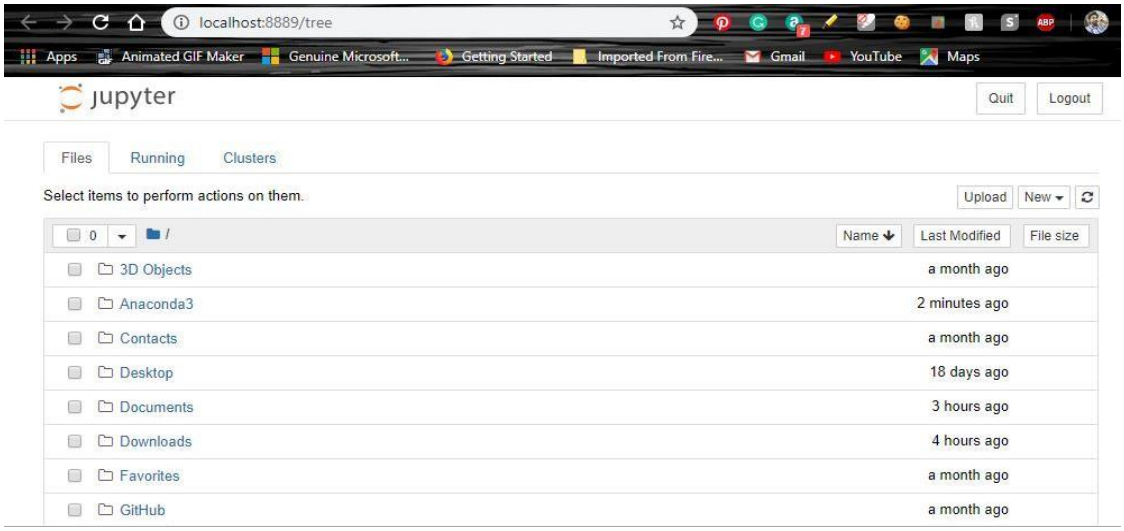


Figure 8: Jupyter Notebook Interface

3.2 Methodology for Analysis of COVID-19 Trend

Detailed methodology followed in this thesis is shown in Figure 9. Regression techniques have been selected for studying the trend and prediction [43] of COVID-19 in Pakistan, selected severely affected country and comparatively less affected country.

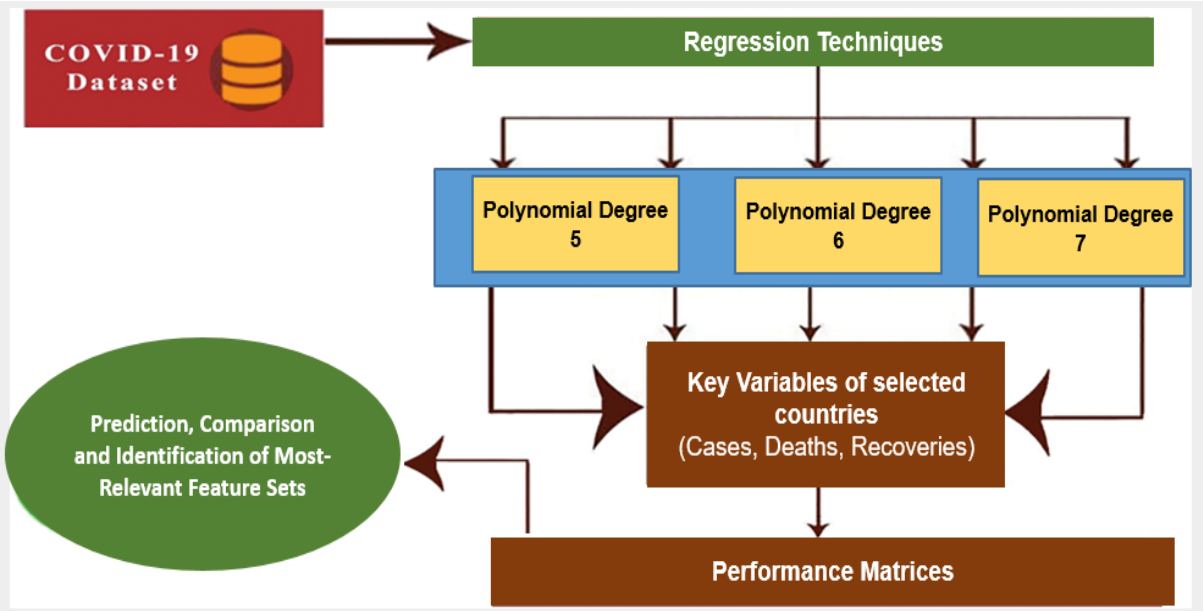


Figure 9: Methodology for COVID-19 Data Analysis

3.2.1 Selection of Countries

The major country of interest for this research is Pakistan. However, other countries are also selected to study the suitability of ML techniques with geographical area bounds.

Italy has been selected as severely affected COVID-19 country. The reason for selection of Italy is because it faced a deteriorating condition than other countries in Europe. Also, as disease was in its early stages therefore it was detected at quite a later time. Moreover, measures were imposed when a lot of regions were already affected with this disease i.e. 20 regions.

The reason for selection of Australia as comparatively less COVID-19 affected country is that at initial stages of disease a high rise in cases was seen in Australia, although for a very limited time, from July 2020 to Aug 2020. However, with time the country was able to control the disease by forcing COVID-19 measures and SOPs around the country and, in result, they were able to control the situation from becoming worse.

3.3 Data Set

For this thesis COVID-19 data has been collected from online source [44] which updates cumulative data of all countries every day. Pakistan and other countries, selected for the research, data has been collected up to 13 months since identification. Data set for 3 variables confirmed cases, deaths and recoveries has been imported directly in python notebook, its pre-processing has been done to bring it in a form to be used in ML polynomial regression technique. The set of commands to import data for defined time period are shown in Appendix A.

3.3.1 Pre-Processing of Data

As mentioned before, this thesis aims to study daily cases data, therefore, daily cases data for all three variables was imported from online source and resulting values were stored in an array for all three countries. The daily data has been considered as dependent variable and date array has been set as independent variable. All the daily data arrays were reshaped to (385x1) vector.

For modelling, data set has been test and trained using python train_test_split() function [45]. For this study data set has primarily been modelled for all variables using 80-20 train/test split size and for comparison of results and performance parameters, ML technique with 95-5 train/test size has also been performed. fit_transform() function is used to scale and fit the data [45]. After these steps, data is processed and ready to be used in Regression model [46, 47].

3.3.2 Setting of ML Modeling Technique

Polynomial Regression has been chosen as ML technique for this research. It forms a relationship between dependent i.e. cases and independent i.e. dates variables. The polynomial regression equation adds all the variables till n. Inspiration for selection of this technique for this study has been taken from research work. The advantage of using this technique is that different data points may be modelled by changing suitable degree in accordance with variation in data. This depicts that data points actually make relationship between factors present in even the nonlinear dataset.

As, we have seen COVID-19 data is highly nonlinear so by changing degree of selected technique, analysis curve behavior changes which can fit in participating feature sets very well.

3.3.3 Model Performance parameters

In order to check the reliability of the model for prediction results, performance parameters comes in to play. The major parameters which are defining results are Root Mean Square Error (RMSE), Mean Square Error (MAE) and R_square functions.

3.4 Major Feature Sets Affecting COVID-19 Curve

There are various feature sets that affected COVID-19 curve of Pakistan as massive fluctuations have been observed in 2020. The feature sets have been identified and then used to explain the behavior of curve from the prediction techniques are as follows:

1) Environmental Features

- Seasons - Summers and Winters

2) Cultural Features

- Eid-Ul_Fitr
- Eid-Ul_Adha
- Quaid Day/ Christmas
- New Year Eve

3) Government Policies

- Lock Down
- Uplift of Ban from International flights
- Re-Opening of Schools

Chapter 4: Deduction of Results and Feature Sets

This chapter presents practical implementation of methodology and deduction of results.

4.1 Modelling and Prediction Using Polynomial Regression

The data has been modelled using the three variants; selected by method of trial and error after applying all variants (2 to 7) of ML technique on selected countries data.

- Degree 5 polynomial Regression
- Degree 6 Polynomial Regression
- Degree 7 Polynomial Regression

The 3 basic indicators of Pakistan, Italy and Australia have been taken in to account on which all these variants have been applied. The indicators are as follows

- Daily Confirmed cases
- Daily Death cases
- Daily Recoveries

4.1.1 Data Set of Pakistan, Italy and Australia

The data set of Pakistan, Italy and Australia is quite varying from each other. Highest number of cases are in Italy, lowest number of cases are in Australia whereas Pakistan stands a medium country. The comparison of cases in all three countries is shown in Fig 10. Similarly, talking about death rate again from Fig 11 highest deaths can be seen in Italy, then in Pakistan and lastly in Australia. However, Pakistan observed first wave when cases and deaths were quite low in Italy and also condition was stable in Australia. Fig 12 shows comparison of recoveries of all three countries, though Italy was critically affected, its recovery index is also quite high.

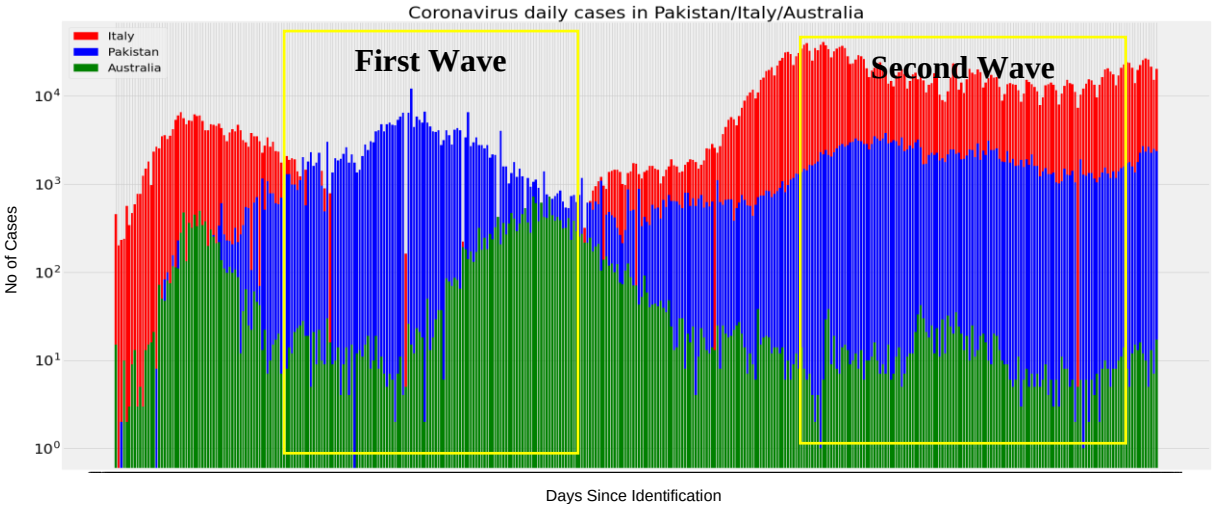


Fig 10. Comparison of COVID-19 Daily Cases in Pakistan/Italy/ Australia

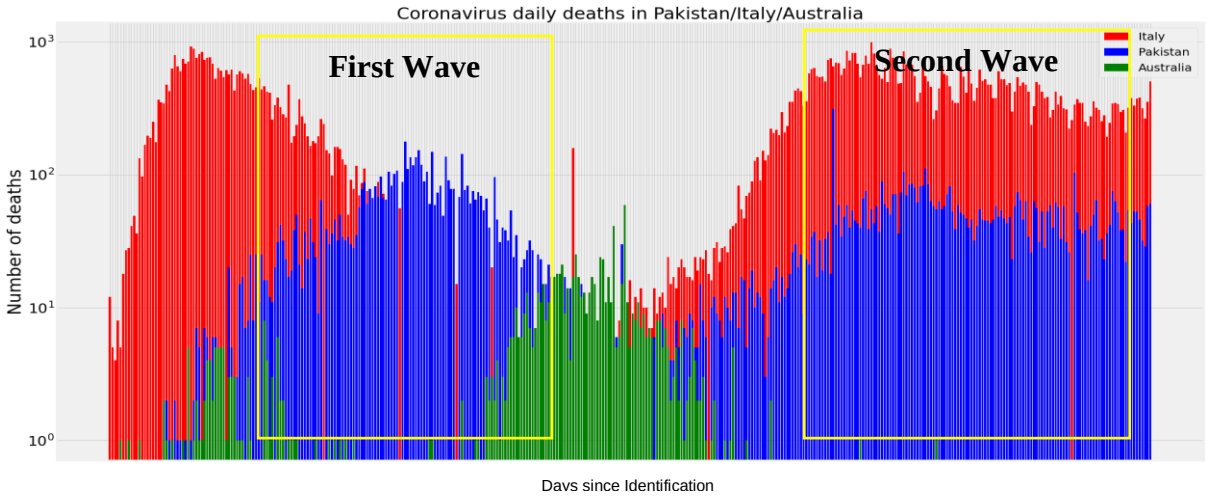


Fig 11. Comparison of COVID-19 Daily Deaths in Pakistan/Italy/ Australia

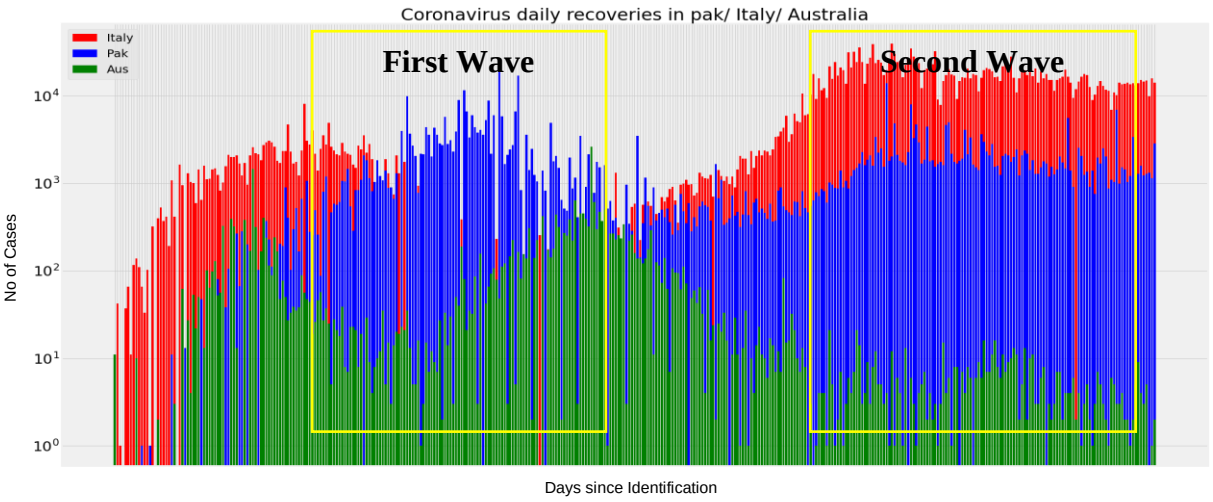


Fig 12. Comparison of COVID-19 Daily Recoveries in Pakistan/Italy/ Australia

4.2 Modelling of Data sets

Selected countries have been modeled using all 3 variants of polynomial regression and then best suited degree has been selected on the basis of performance parameters. In all graphs 'number of cases' are on y-axis and 'days since identification' on x-axis. Also, solid line shows actual trend and dotted line shows predictions. Moreover, all graphs have been generated on 80-20 train/test split size.

4.2.1 Modelling of Pakistan Data Set

Pakistan daily cases, deaths and recoveries data has been modelled using degree 5, 6 and 7 for 13 months since identification of virus.

Cases

Modeling of Pakistan daily cases is shown in Fig 13. Analysis line is best mapped by degree 6 with lowest errors, MAE is 445 and RMSE is 615. Also, R-score is 0.68 which gives around 70% accuracy of predicted results

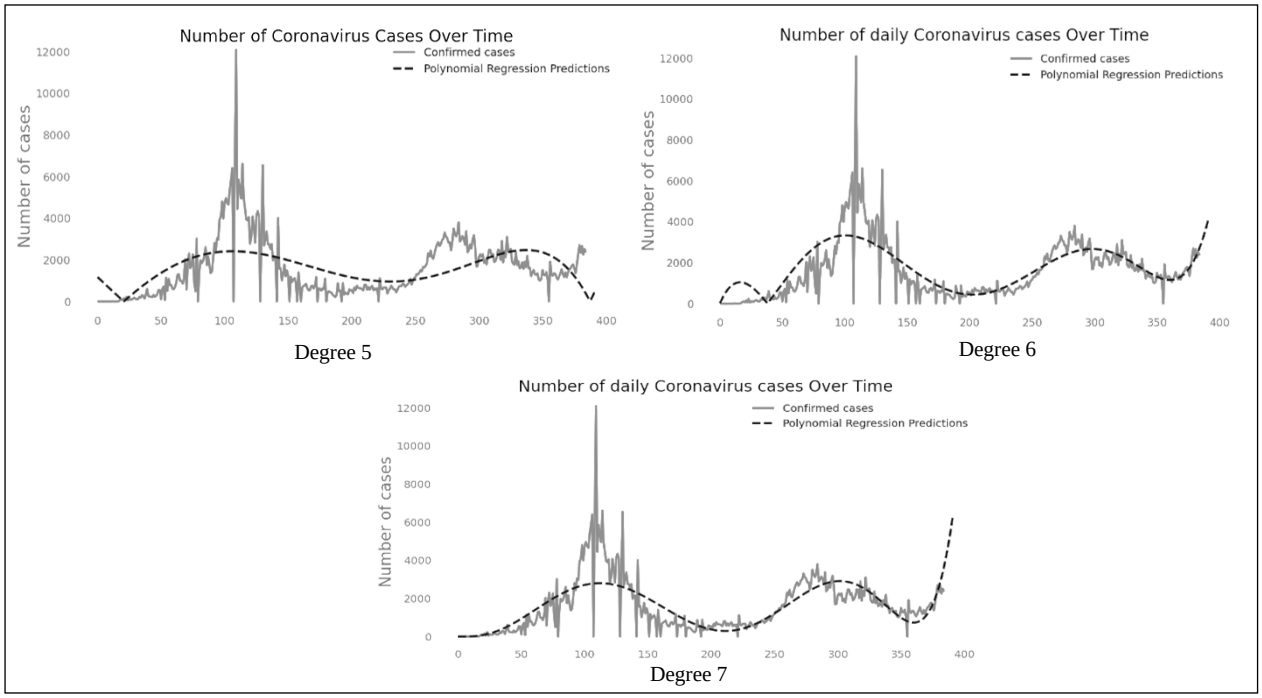


Fig 13. Modelling of Daily Pakistan Cases

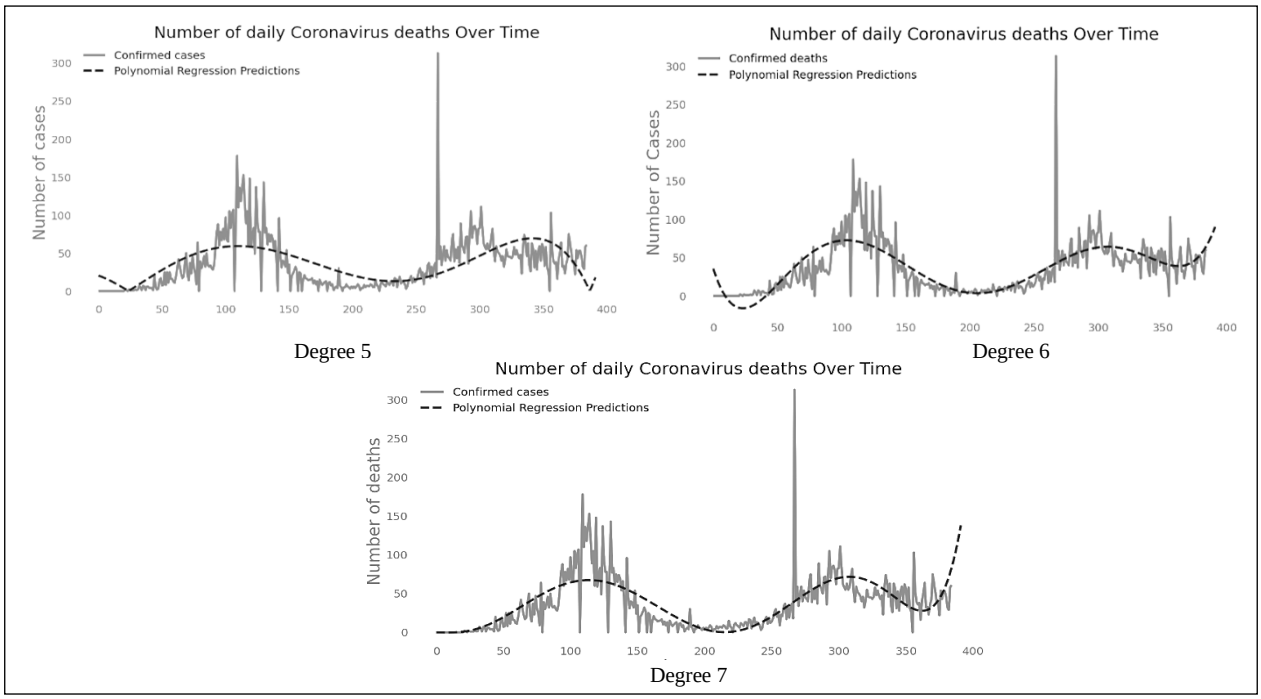


Fig 14. Modelling of Pakistan Daily Death Cases

Deaths

Modeling of Pakistan death cases is shown in Fig 14. Analysis line is again best fitted by degree 6 as MAE is 16 and RMSE is 35. Although RMSE is bit high but R-score is 0.80 which gives 80% accuracy of predicted results, higher than other two variants.

Recoveries

Modeling of Pakistan recoveries data is shown in Fig 15. Efficient results are achieved by degree 7 as its MAE and RMSE are 827 and 1537 respectively. Also, R-score is 0.73 which depicts predictions are around 70% accurate.

4.2.2 Modelling of Australia Data set

Australia Cases, deaths and recoveries have been modelled by using aforementioned 3 variants of polynomial regression.

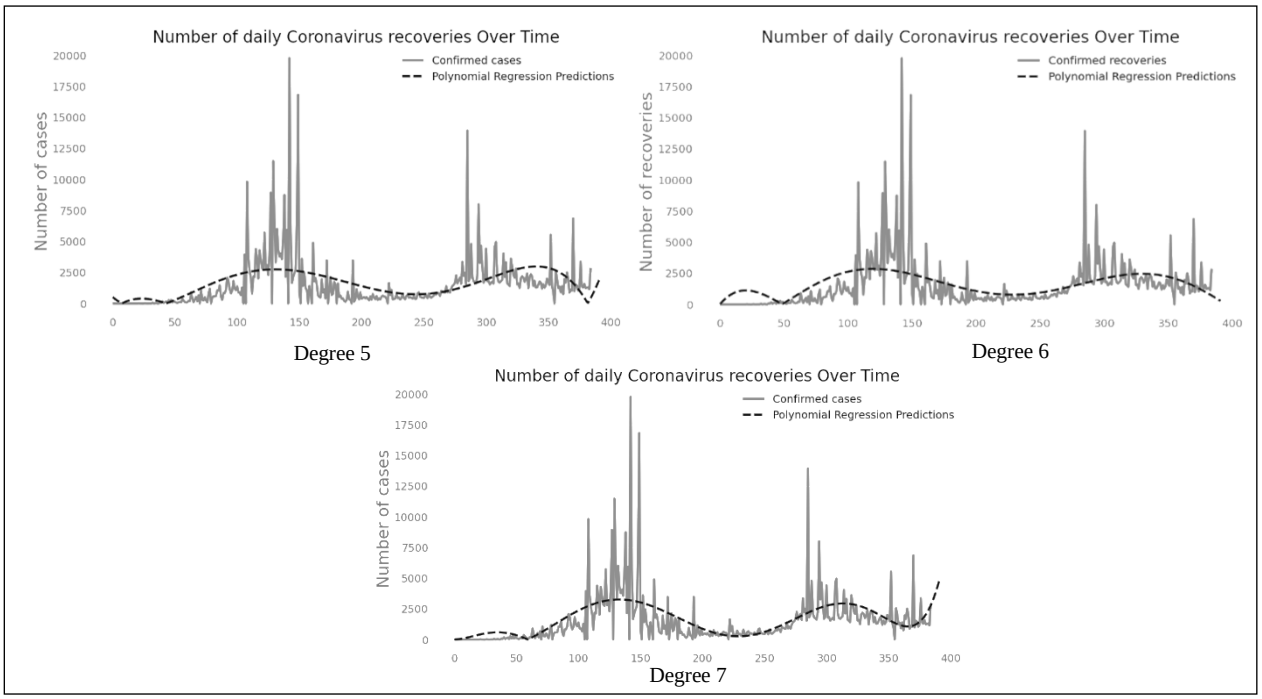


Fig 15. Modelling of Pakistan Daily Recoveries Cases

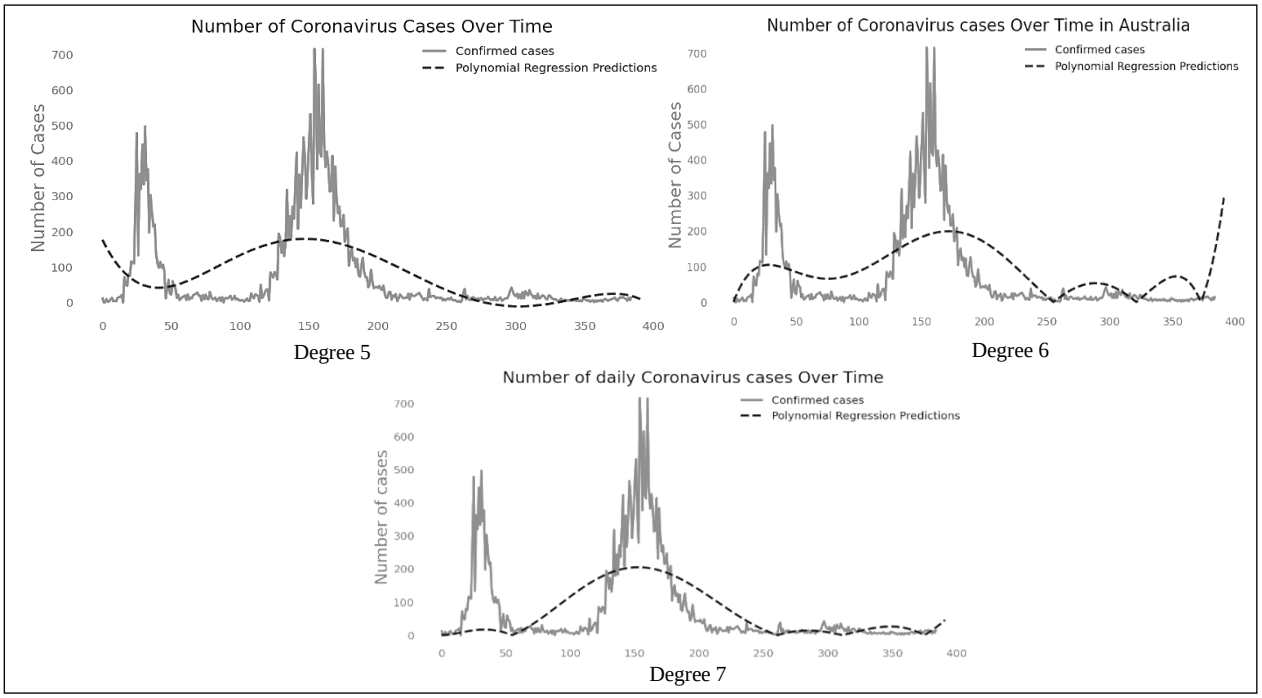


Fig 16. Modelling of Australia Daily Confirmed Cases

Cases

Modelling of Australia cases is shown in Fig 16. Cases are best modelled by degree 5 as MAE is 63 and RMSE is 100. Also, R-score is 0.97 which shows predicted results are around 90% accurate.

Deaths

Fig 17 shows modelling of death cases of Australia. It has been best modelled using degree 5 as MAE and RMSE is 14 and 3 respectively. Also, predicted results are around 80% accurate as R-score comes out to be 0.84.

Recoveries

Modelling of Australia recoveries data is shown in Fig 18. Recoveries data has been best modelled by using degree 5 as MAE and RMSE is 65 and 88. Also, R-score is 0.97 which indicates that accuracy of predicted results is 97%.

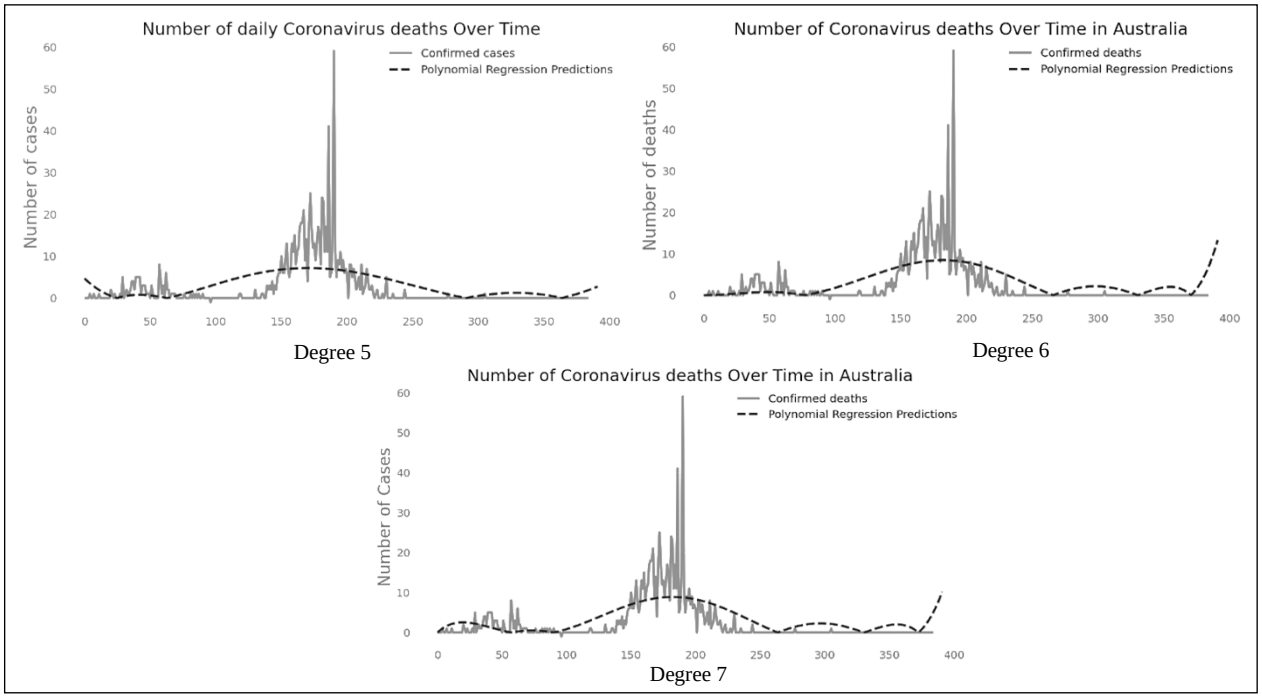


Fig 17. Modelling of Australia Daily Death Cases

4.2.3 Modelling of Italy Data set

Cases

Fig 19 shows modelling of Italy confirmed cases. Its cases are best analyzed by degree 6 as MAE and RMSE is 3680 and 5337 respectively. Also, R-score is 0.65 which depicts accuracy around 70% of predicted cases

Deaths

Italy Death cases shown in Fig 20 have been best modeled by degree 6 as MAE is 121 and RMSE is 146. Also, results are around 60% accurate as R-score is 0.55.

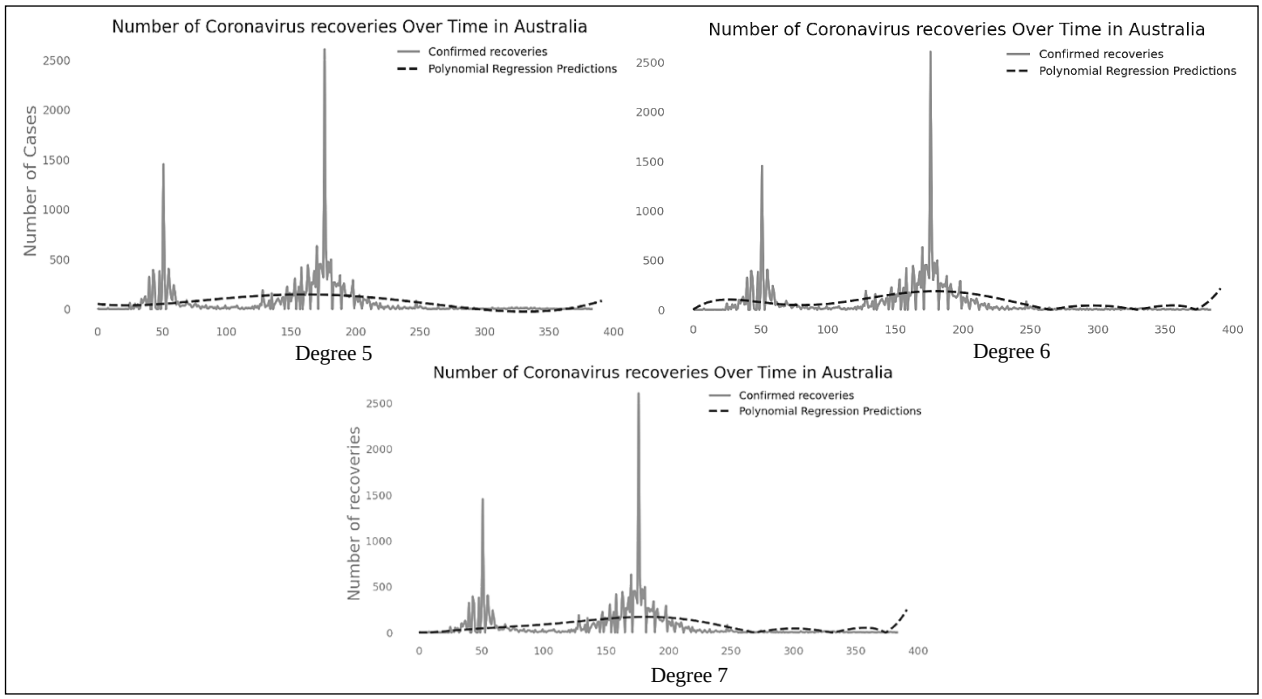


Fig 18. Modelling of Australia Daily Recovered Cases

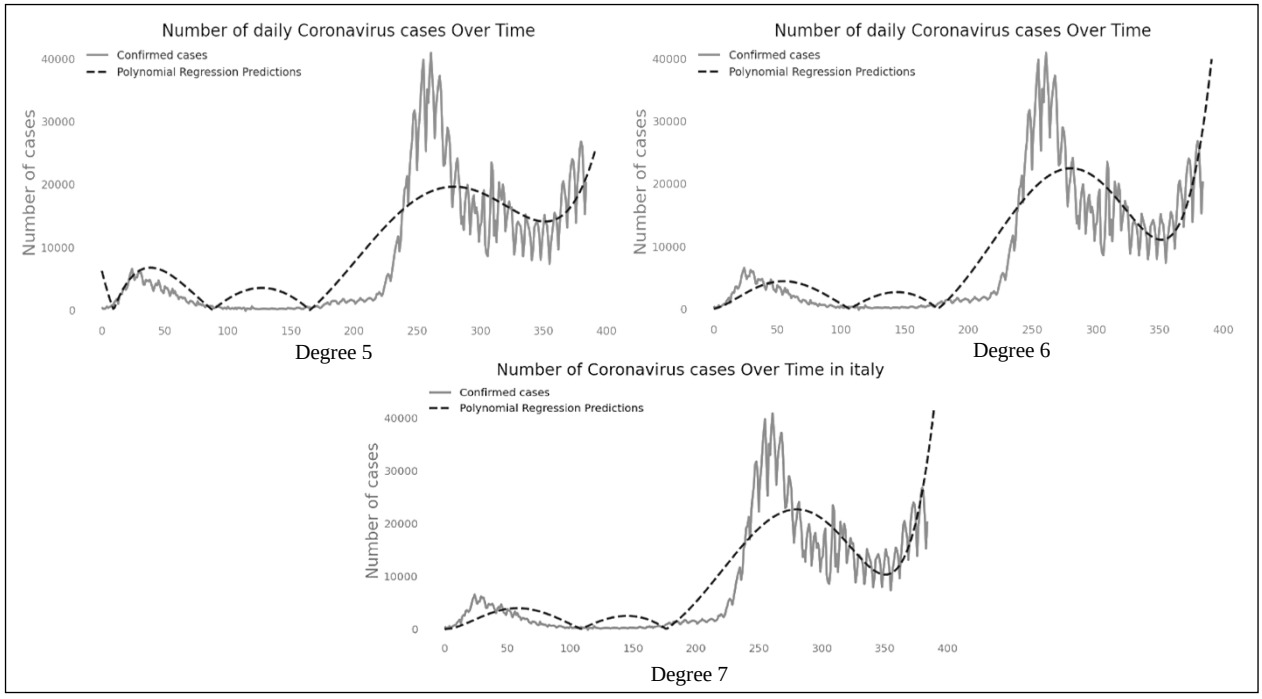


Fig 19. Modelling of Italy Daily Confirmed Cases

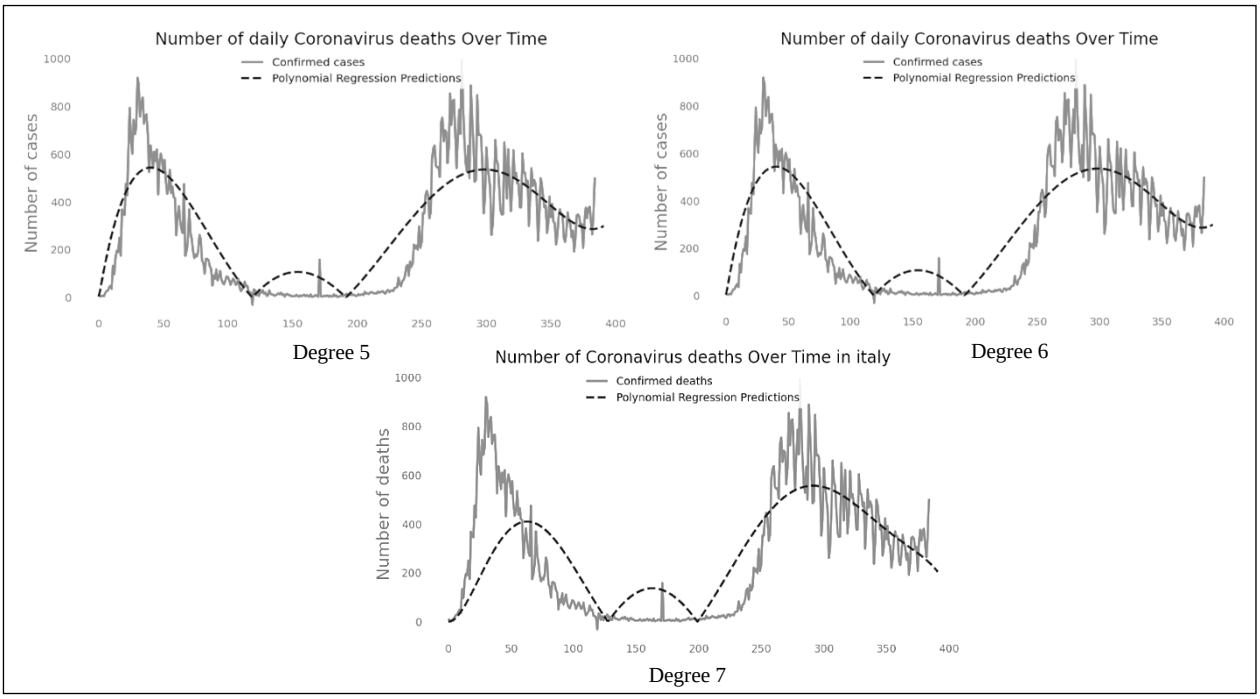


Fig 20. Modelling of Italy Daily Death Cases

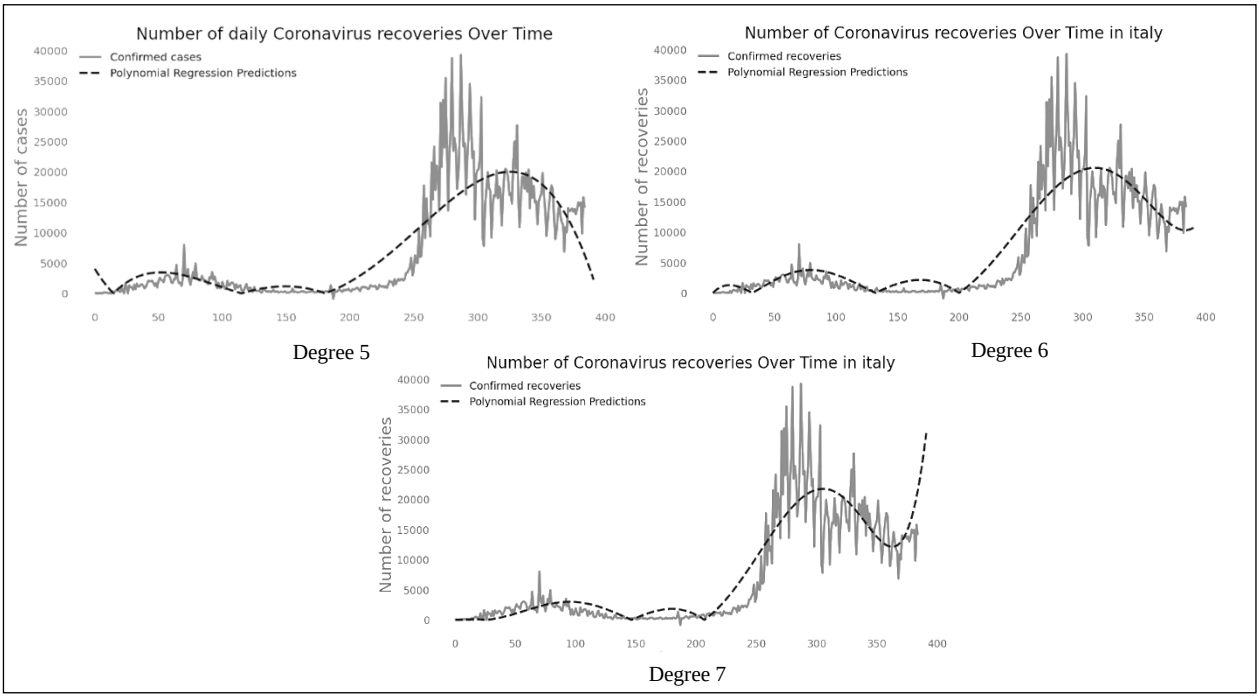


Fig 21. Modelling of Italy Daily Recoveries Cases

Recoveries

Fig 21 shows recoveries cases of Italy, best modelled by degree 7 as MAE is 2450 and RMSE is 4309. Also, predictions are around 70% accurate as R-score is 0.74.

4.3 Prediction of cases

This section includes predicted confirmed cases and death cases, in number, with 80-20 train/test size and with cases predicted by 95-5 train/test size (for results comparison). A comparison of performance parameters of both splits has been done in next chapter.

4.3.1 Pakistan

Prediction results of cases from best suited variant is shown in fig 22 and 23 i.e. Degree 6. The predictions from both splits are close to actual cases i.e around 70%, however, results with 80-20 split have less difference with the actual cases.

	Date	Predicted number of cases in pakistan
0	03/17/2021	2937.0
1	03/18/2021	3112.0
2	03/19/2021	3298.0
3	03/20/2021	3495.0
4	03/21/2021	3704.0
5	03/22/2021	3926.0
6	03/23/2021	4160.0

	Date	Predicted number of Confirmed deaths in pakistan
0	03/17/2021	50.0
1	03/18/2021	51.0
2	03/19/2021	53.0
3	03/20/2021	55.0
4	03/21/2021	57.0
5	03/22/2021	59.0
6	03/23/2021	62.0

Fig 22: Predicted Cases / Deaths in Pakistan with 80-20 Split

Similarly, for death cases with 80-20 split prediction results are more close to actual cases, 80% accurate. However, with 95-5 split, difference with actual cases is quite high reducing the accuracy to around 60%.

	Date	Predicted number of Confirmed cases in pakistan
0	03/17/2021	2713.0
1	03/18/2021	2875.0
2	03/19/2021	3048.0
3	03/20/2021	3232.0
4	03/21/2021	3428.0
5	03/22/2021	3635.0
6	03/23/2021	3855.0

	Date	Predicted number of Confirmed deaths in pakistan
0	03/17/2021	68.0
1	03/18/2021	71.0
2	03/19/2021	74.0
3	03/20/2021	78.0
4	03/21/2021	82.0
5	03/22/2021	86.0
6	03/23/2021	90.0

Fig 23: Predicted Cases / Deaths in Pakistan with 95-5 Split

4.3.2 Australia

Prediction results of cases with both splits, from best modelling i.e. degree 5, are shown in Fig 24 and 25. The predicted confirmed cases are around 90% accurate with 80-20 split. However, with 95-5 split accuracy of confirmed cases is 95% as the predictions are closer to actual cases observed. Similarly, for deaths results with both splits are 80% accurate. The reason for high accuracy is that variations in data set of Australia is very less as disease was controlled in its early stages, therefore, data points are more accurately mapped by the technique variant.

	Date	Predicted number of Confirmed cases
0	03/17/2021	3.0
1	03/18/2021	5.0
2	03/19/2021	8.0
3	03/20/2021	10.0
4	03/21/2021	13.0
5	03/22/2021	16.0
6	03/23/2021	19.0

	Date	Predicted number of Confirmed deaths in Australia
0	03/17/2021	3.0
1	03/18/2021	3.0
2	03/19/2021	3.0
3	03/20/2021	3.0
4	03/21/2021	4.0
5	03/22/2021	4.0
6	03/23/2021	4.0

Fig 24: Predicted Number of Cases / Deaths in Australia with 80-20 Split

	Date	Predicted number of Confirmed cases in Australia
0	03/17/2021	17.0
1	03/18/2021	16.0
2	03/19/2021	15.0
3	03/20/2021	13.0
4	03/21/2021	12.0
5	03/22/2021	10.0
6	03/23/2021	9.0

	Date	Predicted number of Confirmed deaths in Australia
0	03/17/2021	2.0
1	03/18/2021	2.0
2	03/19/2021	3.0
3	03/20/2021	3.0
4	03/21/2021	3.0
5	03/22/2021	3.0
6	03/23/2021	3.0

Fig 25: Predicted Number of Cases / Deaths in Australia with 95-5 Split

4.3.3 Italy

From 80-20 split, prediction results of cases are shown in fig 26 with best modeled variant i.e. degree 6. The confirmed cases and deaths are quite high, still predictions are 60% accurate and close to actual cases observed by Italy. For 95-5 split, prediction results for cases and deaths of Italy from best modelling degree 5 and degree 7 are shown in Fig 27 respectively. Its predictions are 50% closer to observed cases. Overall, accuracy for Italy cases is less because trend is highly fluctuating due to high number of cases or their might be an error in reporting of data.

	Date	Predicted number of Confirmed cases
0	03/17/2021	30076.0
1	03/18/2021	31524.0
2	03/19/2021	33044.0
3	03/20/2021	34638.0
4	03/21/2021	36307.0
5	03/22/2021	38054.0
6	03/23/2021	39881.0

	Date	Predicted number of Confirmed deaths in Italy
0	03/17/2021	303.0
1	03/18/2021	303.0
2	03/19/2021	303.0
3	03/20/2021	304.0
4	03/21/2021	305.0
5	03/22/2021	307.0
6	03/23/2021	309.0

Fig 26: Predicted Number of Cases / Deaths in Italy with 80-20 Split

	Date	Predicted number of Confirmed cases in Italy
0	03/17/2021	12783.0
1	03/18/2021	12986.0
2	03/19/2021	13205.0
3	03/20/2021	13440.0
4	03/21/2021	13694.0
5	03/22/2021	13964.0
6	03/23/2021	14254.0

	Date	Predicted number of Confirmed deaths in Italy
0	03/17/2021	238.0
1	03/18/2021	233.0
2	03/19/2021	228.0
3	03/20/2021	222.0
4	03/21/2021	217.0
5	03/22/2021	211.0
6	03/23/2021	204.0

Fig 27: Predicted Number of Cases / Deaths in Italy with 95-5 Split

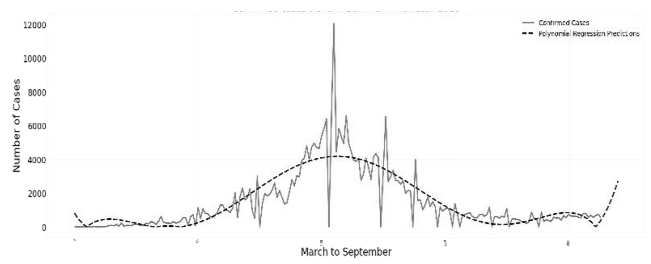
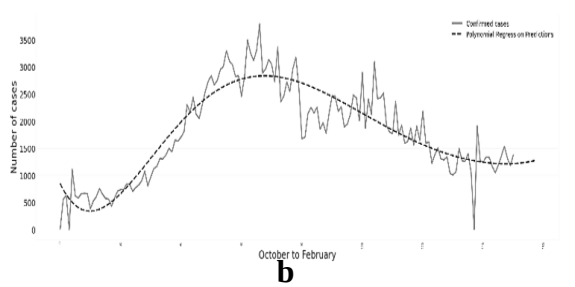
4.4 Evaluation of Feature Sets in Pakistan Curve

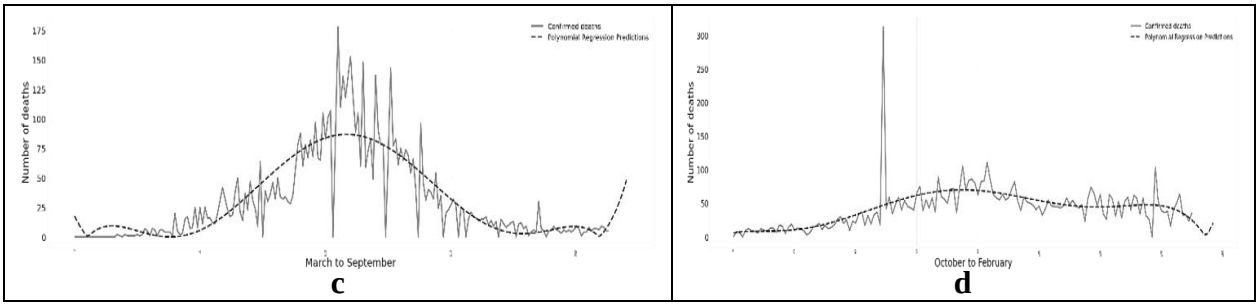
There are many environmental, cultural and government policies features which have played their role in increasing COVID-19 cases in Pakistan. It is important to learn which feature has what impact on COVID cases. Since, 1 year of disease has been

passed in Pakistan but still the cases and deaths are high and the end of this disease seems nowhere hence, it may be assumed that same features will repeat themselves in the upcoming days and might possible they follow the same historic trend. In all the graphs below, solid line depicts actual cases and dotted line shows predicted results.

4.4.1. Environmental Factors (Seasons)

For the COVID-19 curve two different peaks have been seen in summers and winters which lead to a fact that weather has impact on disease cases. Therefore, trend has been studied in both seasons using derived ML technique on all selected factors. February 2020 to September 2020 has been considered for summers and October 2020 to February 2021 for winters. From the simulation results shown in Fig 28, it can be inferred that seasons is not a major participating feature however it has minimal effect. The deaths and cases are observed quite high in summers as compared in winters the trend is comparatively smooth which shows cases were there but severity was less as compared to summers. First COVID-19 wave hit the country in summers; second wave in winters. In the similar times high cases have been observed and now predictions have indicated a third wave will hit the country in near future, which is currently ongoing now. Moreover, which changing climate masses change their habits for indoor and outdoor activities which results in to variation in cases accordingly.

Summer (Feb to September)	Winter (October to February)
	
a	**b**



c d

Fig 28: Environmental Features; a) Summer Cases b) Winter Cases c) Summer Deaths d) Winter Deaths

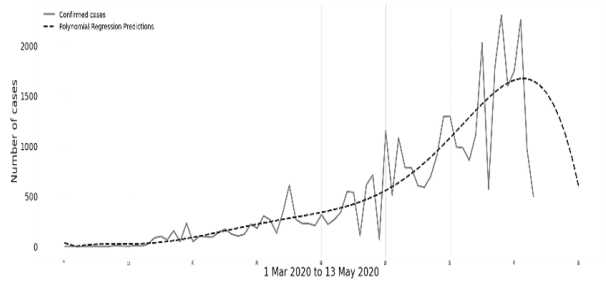
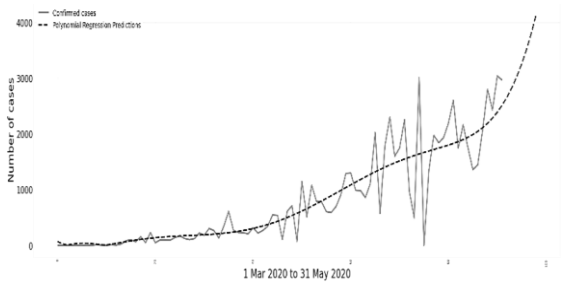
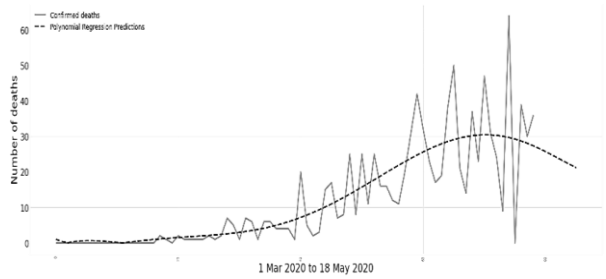
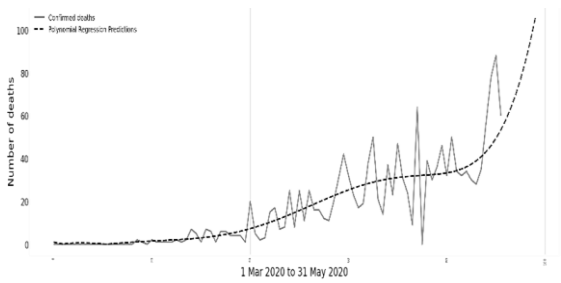
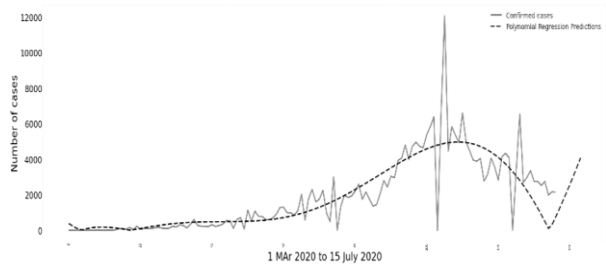
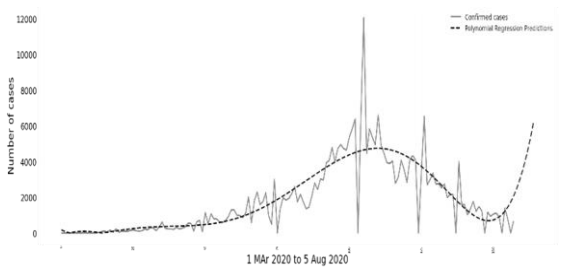
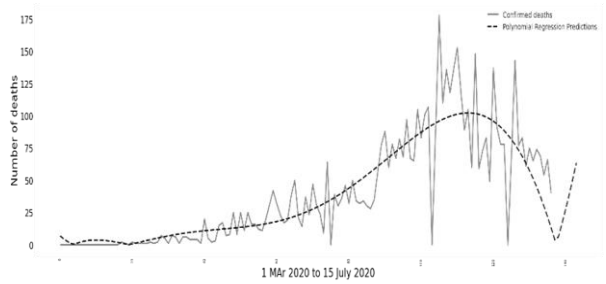
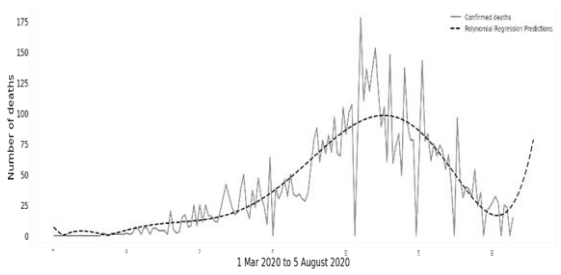
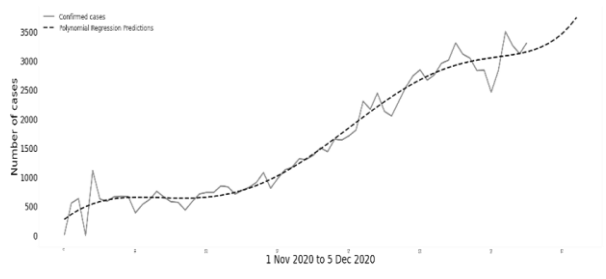
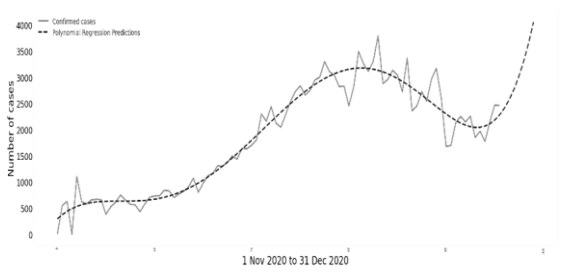
4.4.2 Cultural Features

Cultural factors Eid-ul-Fitr, Eid-ul-adha, Quaid Day, Christmas and New Year Eve have been considered. Analysis has highlighted that few factors have very great impact on COVID-19 cases and few have very slight effect as shown in Fig 29.

Firstly, Eid-ul-Fitr festival was observed with great zeal and zest even during COVID-19 and lock down was also eased out. The analysis of cases 'before said festival' and 'after said festival' shows that cases showed a drastic increase after the festival. It is one of the major factor participated in rise in cases as Shopping areas ware opened and overwhelming number of people found their way to markets. Prediction results before said event shows that cases would have been much less and trend followed would have been linear if SOPs were not eased out.

Secondly, talking about Eid-ul-Adha after seeing the results it has slight impact on cases. Although, cases rose to quite a number as compared to cases going on before festival but still it is not a major participating factor as compared to EID-UL-Fitr festival as no drastic change was observed.

On 25 Dec festival cases again rose to quite a number on and after the festival. Same results also shows rise in cases on New Year eve as the people gathered together to celebrate these events which resulted in spread of disease.

Cases Before Festival	Cases after Festival
a. Eid-UL-Fitr	
	
	
b. Eid-UL-Adha	
	
	
c. 25 Dec Festival (Quaids Day & Christmis and New year Eve)	
	

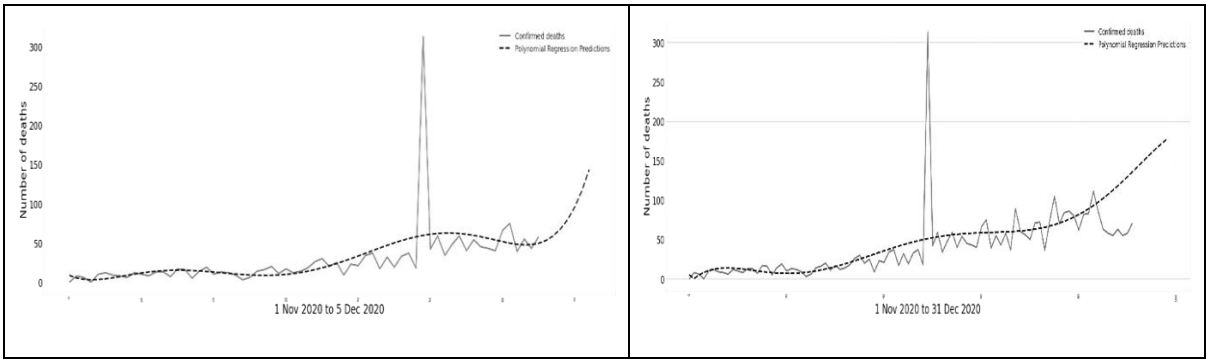


Fig 29: Impact of Cultural Features; a) Eid-ul-Fitr Cases b) Eid-ul-Adha Cases c) Christmas and New Year Eve Cases

4.4.3 Government Policies

Lock Down

An influencing factor in disease curve is Lock down which was practiced on 21 March 2020 but uplifted on 9 May 2020. Analysis have shown that cases and deaths have increased almost twice after uplift of lock down. Also prediction analysis shows if lock down had been carried on, the results would have been much lower, however, trend after lock down was quite drastic as shown in Fig 30.

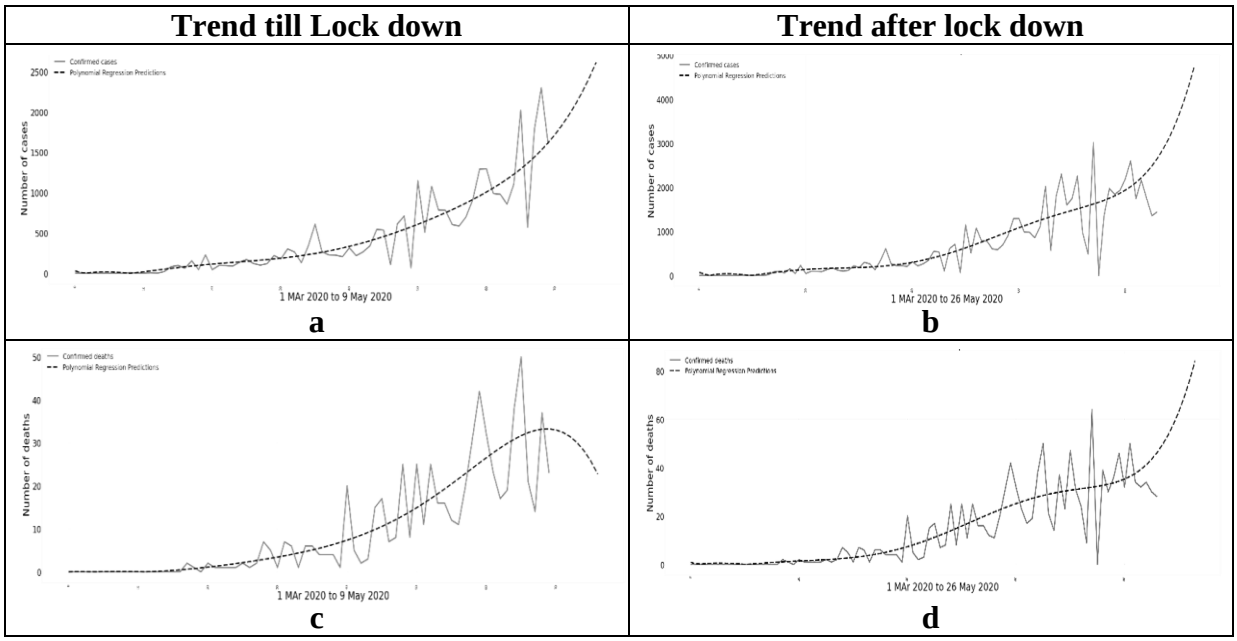


Fig 30: Impact of Lockdown; a) Cases Before b) Cases After c) Deaths before d) Deaths After

Uplift of Ban on International Flights

It could have been devastating factor, however right decision was taken at right time. International flights were banned on 21 March till 7 Aug when the situation was quite stable in foreign countries. Hence, analysis of said feature has shown that it don't have much larger impact on corona cases but only a slight effect as shown in Fig 31, which don't had any major consequences.

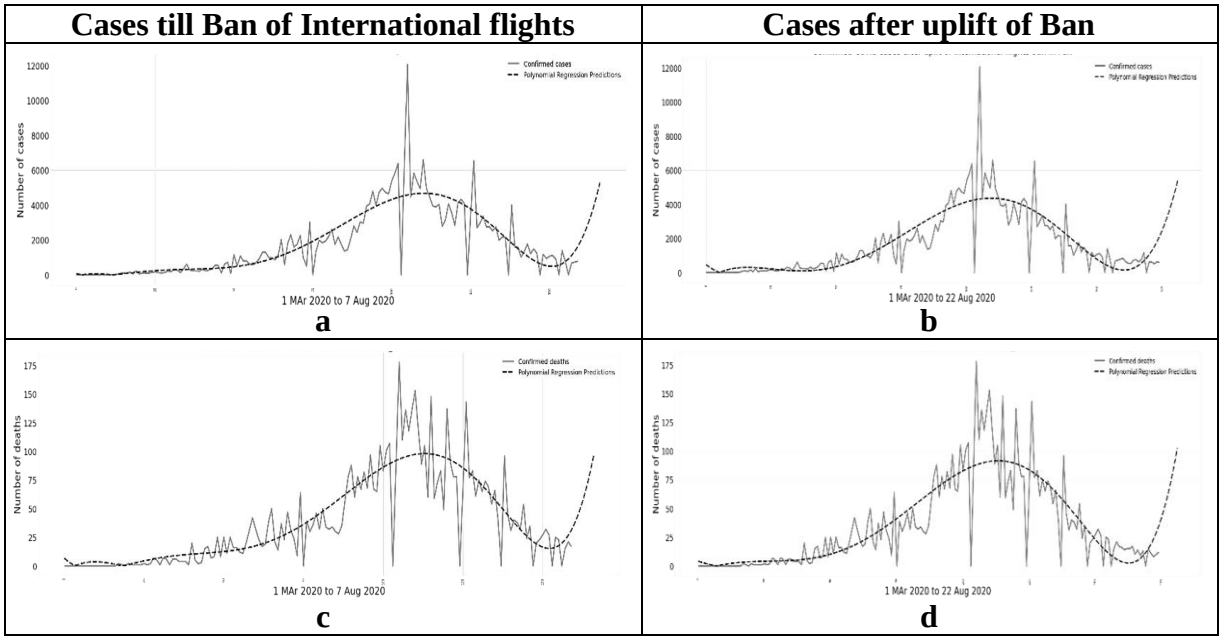


Fig 31: Impact of Uplift of Ban from International Flights; a) Cases Before b) Cases After c) Deaths Before d) Deaths After

Re-opening of Schools

Education sector and student career has majorly been affected from pandemic not in terms of health but also in economic means. The analysis shows that it is not a major participating factor in rise of cases because schools were closed at right time from 13 Mar to 15 Sep. However, after the opening of schools slight rise has been observed but no major fluctuations are seen as shown in Fig 32.

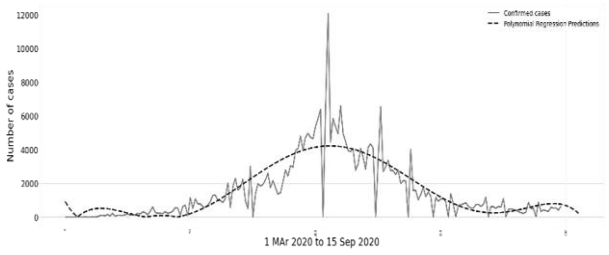
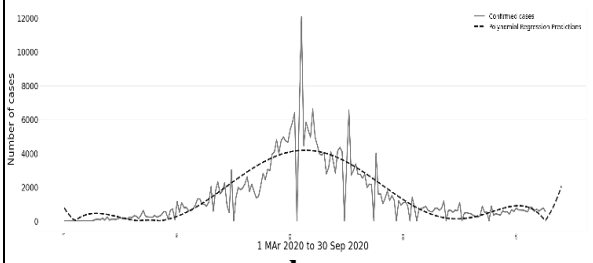
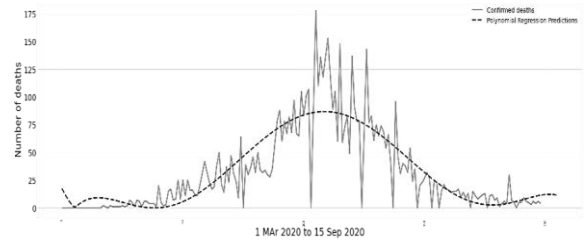
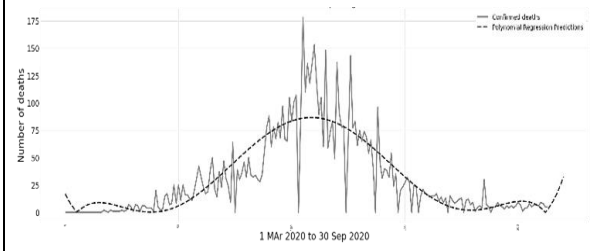
Cases till closure of schools	Cases after opening of schools
 **a**	 **b**
 **c**	 **d**

Fig 32: Impact of Re-opening of Schools; a) Cases Before b) Cases After c) Deaths Before d) Deaths After

Chapter 5: Analysis and Proposed Framework

5.1 Modelling of Selected Countries

The research has been performed on daily data instead of cumulative cases to get deep analysis of variation of cases in Pakistan, Italy and Australia. Australia has been selected as less affected COVID-19 country among all as a consequence of cases it witnessed in July 2020 to Aug 2020 were high, although for very short period, however, eventually it controlled the spread extremely well till now by implementing basic COVID SOPs. Italy has been selected as severely affected COVID country owing to its worst condition in comparison to other European countries. They detected the virus at a very later stage and so were the SOPs implemented meanwhile disease had affected all 20 regions of the country.

ML technique with three different variants has been applied on the data of all three countries with 80-20 train/test split size and on the basis of performance parameters best modelling has been selected. We have tested datasets on degree 5, 6 and 7 because we found through trial and error that best estimations are appearing within this range of variants. Since, accuracy is inversely proportional to error, accordingly lower the error higher will be accuracy. Using the aforementioned fact and prediction graphs/ figures, we have deduced our results. The models have been trained on different matrices and analysis has been done along with prediction of results.

The results have also been performed on 95-5 train/test split size for comparison of performance parameters and results. With this split, the best estimations have again been found in the range of degree 5, 6 and 7. Accuracy with respect to all variables in detail has already been discussed a/w results in the previous chapter. Results are

almost similar to the results with 80-20 split with difference of 10-15% in results. Comparison of performance parameters for both splits are shown in Table IV and V respectively for all variants and all variables.

Variation in COVID-19 cases of Australia is lower, deaths due to this disease have been NIL since 2020, however, cases below 30 per day could be observed. Using selected technique we fitted the polynomials and determined that best variant for Australia data set is degree 5 with minimum error and high accuracy for all variables for both splits with a difference of 5-10% in results. Also, table IV and V shows that MAE and RMSE for degree 6 is higher than degree 5 and for degree 7 errors are highest for both splits. Therefore, curve of Australia for all factors has further been studied using the same derived variant. Also, the graphs for well suited variant (degree 5) derived in previous chapter analyze the trend of cases and deaths in accordance with ongoing situation in Australia. Prediction curve is upside down which exhibits cases will not increase in future and situation will be under control in coming future.

Dataset of Italy has highest variation in terms of cases, deaths and recoveries. Italy data set has also been modelled using all 3 variants and 80-20 train/test split; it has been developed that degree 6 is best suited for confirmed cases/ deaths and degree 7 for recoveries with highest accuracy and minimum errors as shown in Table IV. However, with 95-5 train/test split size degree 7 is best suited for deaths and recoveries and degree 5 for confirmed cases with minimum error as shown in Table V. Moreover, from graphs generated for Italy with fitted variant it has been analyzed that Italy cases will grow in future as predicted line is upwards, however, a decline in rate may be observed to some extent as compared to 2020 cases.

Considering Pakistan data set, after application of all three variants it was discovered that degree 6 polynomial is best fitted for confirmed cases and deaths. On the contrary, recoveries have been analyzed by degree 7 polynomial regression. The best modelled variants are same for both split size, however, accuracy and performance parameters are varying as shown in table IV and V, with difference of 10-15% results in both splits. Lower order polynomials are not suitable reason being errors are high and also the fluctuating curve's varying data points are not fitted by lower order trend line. Similarly, degree 7 is also conferring high errors on Pakistan cases and deaths as shown in Table IV. In some cases, MAE and RMSE of variables are observed to be same, so in such cases, best variant has been considered by taking in account accuracy and deduced predicted results. Pakistan best fitted prediction curve reveals that all three indicator cases will increase in near future as the predicted trend line is upwards. It indicates another upcoming severe corona virus wave in Pakistan, which is ongoing.

Also, it has been observed from performance parameters that although MAE and RMSE is, in some cases, lower with 80-20 split and, in some cases, it is higher as compared to 95-5 split. Detailed results are shown in table IV and V, also, performance parameters of best fitted variants have been highlighted. However, predicted results are closer to actual cases with 80-20 split giving high accuracy.

	Daily cases						Daily deaths						Daily recoveries					
Deg	Pakistan		Italy		Australia		Pakistan		Italy		Australia		Pakistan		Italy		Australia	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
5	834	1057	4301	5961	**63**	**100**	18	24	121.6	147	**2.6**	**3.8**	907	1304	3304	4938	**65**	**88.3**
6	**446**	**616**	**3680**	**5337**	72	96	**16**	**35**	**121.5**	**146.2**	2.7	3.9	857	1209	2858	4469	67	97
7	593	789	3704	5331	63	107	17	24	128	159	2.8	4.0	**827**	**1537**	**2451**	**4310**	65	89.3

Table IV: Performance Parameters of Pakistan/Italy/Australia Modelling, 80-20 Split

	Daily cases						Daily deaths						Daily recoveries					
Deg	Pakistan		Italy		Australia		Pakistan		Italy		Australia		Pakistan		Italy		Australia	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
5	12794	14842	**6706**	**7563**	**11**	**12**	210	242	387	391	**2**	**2**	5416	5742	14072	14502	**55**	**55**
6	**259**	**231**	9567	11308	162	162	**9**	**13**	396	440	4	4	841	1128	5631	5861	107	107
7	1782	1782	16651	18951	177	189	21	21	**357**	**390**	5	5	**248**	**248**	**3257**	**4236**	115	115

Table V: Performance Parameters of Pakistan/Italy/Australia Modelling, 95-5 Split

A comparison of actual and predicted cases from both splits are shown in Table VI for Pakistan data which revealed that predictions performed by derived variant are reliable as accuracy is prominently high i.e. Cases predicted on 17 March 2021 are 2937 with 80-20 split and 2758 with 95-5 split; observed confirmed cases were 3946 which reveals accuracy of 85% and 79% respectively. Similarly, overall results have been predicted for future dates with accuracy of 70% and above. The difference in %age accuracy among dates is due to changing variations in the curve.

Date	Cases Predicted in Train/ Test Size		Actual Cases	Difference in Train/ Test size		% Within Actual Observation in Train/ Test Size	
	80-20 %	95-5%		80-20%	95-5%	80-20%	95-5%
17 Mar 2021	2937	2758	3495	558	737	15.96	21.09
18 Mar 2021	3112	2926	3449	337	523	9.77	15.16
19 Mar 2021	3298	3105	3876	578	771	14.91	19.89
20 Mar 2021	3495	3295	3667	172	372	4.69	10.14
21 Mar 2021	3704	3497	3889	185	392	4.75	10.08
22 Mar 2021	3926	3711	3270	-656	-441	20.06	13.49
23 Mar 2021	4160	3937	3301	-859	-636	26.02	19.27
24 Mar 2021	4390	4176	3946	-444	-230	11.25	5.83
25 Mar 2021	4583	4428	4368	-215	-60	4.92	1.37
26 Mar 2021	4690	4695	4468	-222	-227	4.97	5.08
27 Mar 2021	4931	4975	4767	-164	-208	3.44	4.36

Table VI: Comparison of Predicted and Observed COVID-19 Cases in Pakistan

5.2 Significance of Features sets

In the comparatively lesser affected country, Australia cases were controlled due to enforcement of anti-COVID-19 measures at the detection of disease and afterwards they never eased it until control of disease. The international flights were restricted, strict whole day lock down was imposed and online schooling was introduced. However, one of the reasons for control of cases could be high technology available with them for online shopping, businesses and schooling. Moreover, they celebrated the festivals with strict distancing measures and participation of controlled number of people in Christmas and New Year celebrations.

Italy had a devastating situation due to the reason that disease was identified very late when it was almost spread in the whole country. Therefore, COVID-19 protocols were not implemented timely which lead to uncontrolled situation. Italy also never took the disease seriously as compared to other European countries. In Feb 2020 Champions League match was observed by massive majority. Moreover, It is a country of large

area so population is more and it has quite lot of aged people more than 65 years which is one of the major factor of disease spread. First wave was observed in mid summers and followed by second wave at start of winters.

Features affecting Pakistan's COVID-19 curve have been presented in the previous chapter in detail. It has been observed from prediction results that some factors have significant impact on disease cases than otherwise. The features have been marked on Pakistan curve which indicates peaks observed in the trend when some feature has critically affected the country. On the contrary, troughs have been observed when features don't have positive role in increasing the cases as shown in Fig 31. Considering cultural factors, Eid-Ul-Fitr, christmas and new year eve have positive role in increasing cases, however, Eid-Ul-Adha don't have major impact on cases. Similarly, uplift of lockdown increase cases to a high extent but opening of schools and international flights are not major features in increasing disease cases.

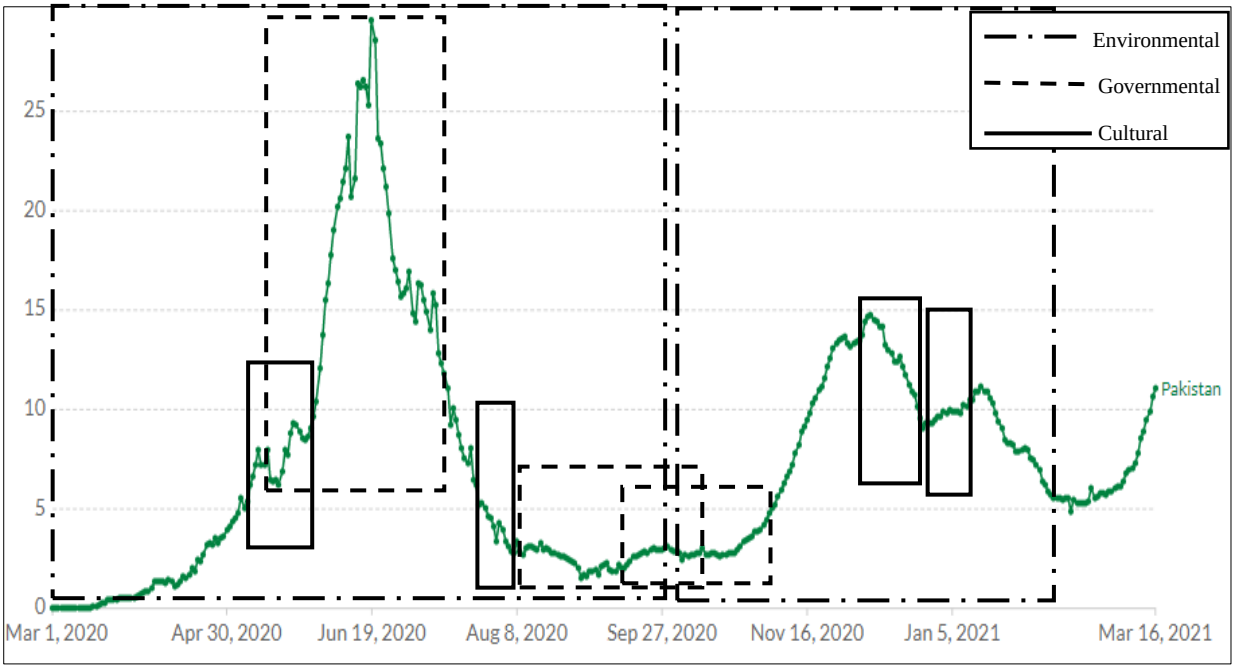


Fig 33: Feature Sets Affecting COVID-19 Curve

5.3 A Proposed Framework

From the methodology which has been implemented in this thesis and the analysis it has been deduced that feature sets play a vital role in not only determining the degree of polynomial regression ML technique but also in increase and decline of the prediction of positive cases, an indicator in the current problem set. Moreover, changing the variant of selected regression technique also changes the prediction trend of all indicators by increasing or decreasing the value of decisive performance parameters which are MAE, RMSE and R_Square, which then further defines the best prediction results for future.

From above discussion it has been understood that to have an automatic mechanism of selecting the best suitable degree/variant of the polynomial regression technique for a particular country data on the basis of key factors would be really helpful in specific research. Real life conditions are different in different environments and due to the variance in these features, prediction curve will be varying for different countries. Therefore, a weight number needs to be assigned to these feature sets, some of which are already identified in thesis, on the basis of effects they have on prediction curve in the past. Number of feature sets give rise to fluctuations in data therefore, variations affect the degree of polynomial due to changing dataset.

In light of these observations we intend to lay out details of an automatic framework which can take the input data and do the prediction curve generation using polynomial regression technique, on the input dataset, automatically. The details of this proposed framework are given below.

For Automation “loop” has been selected, which will run for set of degrees/variants on particular data and on the basis of lower error values and high accuracy it will select

the variant. However, in traditional ML polynomial regression technique all the degrees have to be tested by method of hit and trial and then after achieving satisfying results

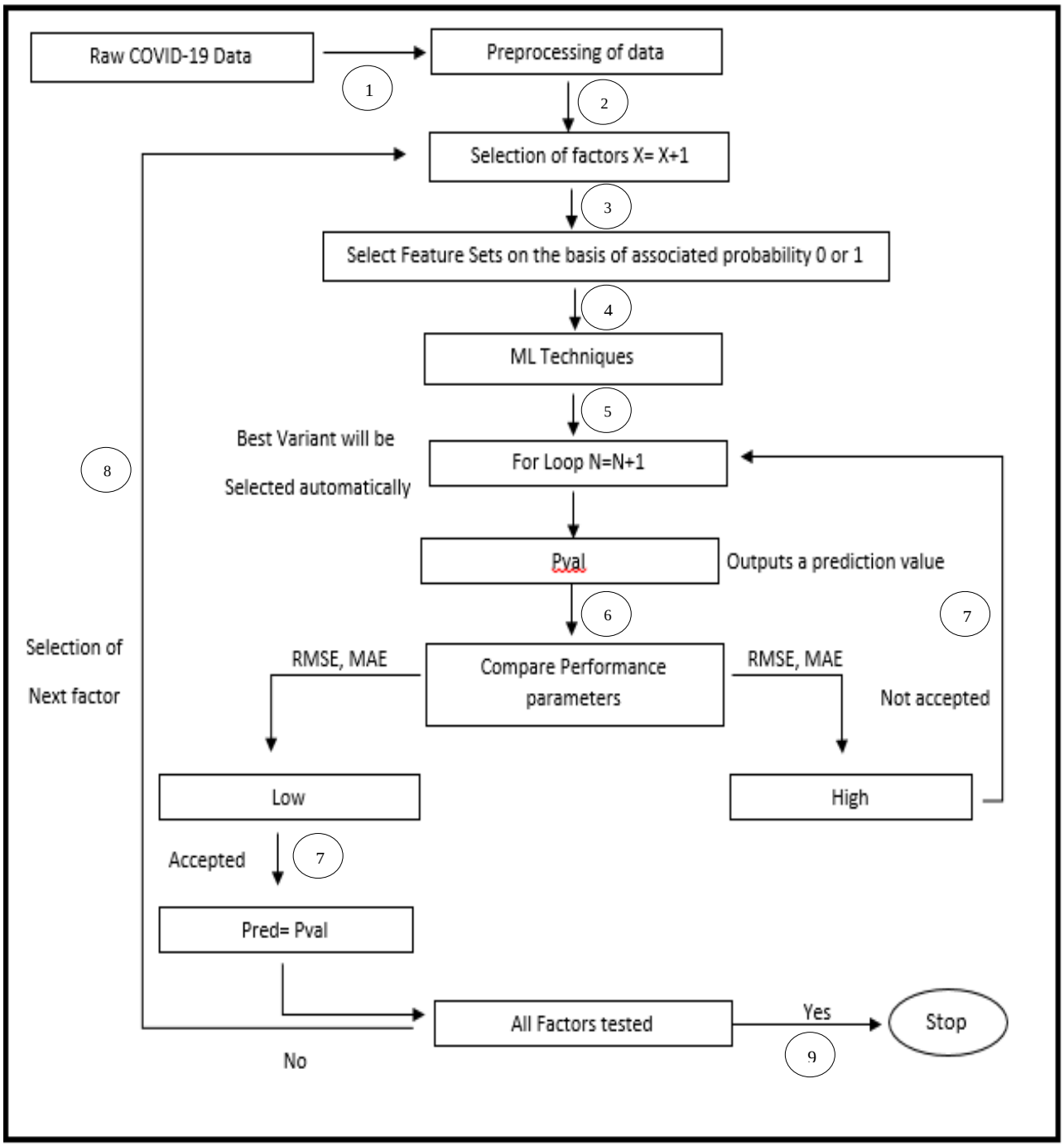


Fig 34: A Proposed Framework

the degree is selected. Also, it give prediction on the basis of latest trend being followed by the historic data without considering the changes that that will occur in the curve due to SOPs implemented, cultural event or seasonal effect.

A Framework has been shown in figure 34, which will automatically select the variant suitable for factor passed and assigns probability value to features as different factors will give a different curve. Also, when feature values will play their role the prediction curve will change as it will no longer be only dependent on the latest trend being followed.

Details of automatic process are as follows

1. Raw data taken from online source will be preprocessed.

2. Any number of indicators will be identified such as cases, deaths etc which will iterate itself automatically.

3. A weight number (probability value from 0 to 1) associated with the feature will be added in the data on the basis of its performance and effect on historic data. F(0,1)=f1+f2+…….+fn

4. Select ML technique

5. Automate the variant associated with the ML technique N=N+1.

6. Compare the output curve on the basis of performance parameters. Pval

7. If error is high and accuracy is low reject it else accept it. N($RMSE_{min}$, MAE_{min})

8. Same process will repeat for other selected indicators. X=X+1

9. Process will terminate after testing of all indicators

Chapter 6: Conclusion and Future Work

6.1 Conclusion

In the thesis variants of polynomial regression have been studied and applied on certain factors of COVID-19 in selected countries, on the basis of which modelling of cases, deaths and recoveries has been done. The variants have been tested using performance parameters, and degree of selected ML technique has then been associated with different countries dataset accordingly. Also, deduction of results and selection of best fitted variant has been tested using 80-20 test/train as well as compared with 95-5 split size.

Evaluations and prediction curves have proved that degree 6 & 7 is suitable for Pakistan COVID-19 trend, however, degree 5 has best modelled Australia curve. Degree 6 & 7 are best suited for Italy with 80-20 and degree 5&7 are best suited for Italy with 95-5 split. The difference in best modelled variant among countries is due to difference in variations in trends. Since, trend of COVID-19 is different in different countries, therefore, degrees are also varying accordingly. Predictions performed for Pakistan trend have proved that after second wave a third COVID-19 wave will hit the country, which is ongoing. Similarly cases will increase in Italy, however in Australia situation will be controlled.

The variability of degrees is due to the fact that distinct feature sets are playing their roles in countries disease curve, leading to rise and fall of disease. Few features have impacted negatively and some had positive impact on the trend. Therefore, the features have been studied and on the basis of them a framework has been proposed which can help automatically predict the suitability of ML technique for varying data.

It has been analyzed that there is still a need to follow COVID-19 measures, although vaccination has been started but still it will take a long time in proving its efficiency. The feature sets studied in the thesis will help the government in framing the policies for future. Also proposed framework if implemented as a tool will lead to effective prediction which will further help health sector to plan strategies early.

6.2 Future Work

In future work on this thesis can be done in following domains

- Modelling of countries other than studied in this thesis can be performed to study in-depth analysis of variants of this disease in various geographical regions
- Various ML techniques can be tested to get close real life modelling of factors
- Identification of Critical feature-sets around the world to analyze and study disease in a generalized term.
- Implementation of proposed framework as a tool

Bibliography

[1] Paules CI, Marston HD, Fauci AS. Coronavirus infections-more than just the common Cold. JAMA 2020;323(8):707–8. doi:10.1001/jama.2020.0757.

[2] Fan Y, Zhao K, Shi Z, Zhou P. Bat coronaviruses in China. Viruses 2019;11(210).

[3] WHO | World Health Organization. https://www.who.int/. Accessed April 14, 2020.

[4] Lan F-Y, Wei C-F, Hsu Y-T, Christiani DC, Kales SN. Work-related COVID-19 transmission in six Asian countries/areas: a follow-up study. PloS One 2020;15(5): e0233588. https://doi.org/10.1371/journal.pone.0233588.

[5] Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner NW, Ke R. The novel Coronavirus 2019-nCoV is highly contagious and more infectious than initially estimated. arXiv, https://doi.org/10.1101/2020.02.07.20021154; 2020.

[6] Zhao S. Estimating the unreported number of novel Coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data-driven modelling analysis of the early outbreak. J Clin Med 2020;9(2):388.

[7] Nishiura H. The extent of transmission of novel Coronavirus in Wuhan, China, 2020. J Clin Med 2020;9(2):330.

[8] Huang C. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395:497–506.

[9] COVID-19 outbreak and probiotics: facts and information from BioGaia | BioGaia.https://www.biogaia.com/other-news/covid-19-outbreak-and-probiotics-facts/. Accessed April 14, 2020

[10] Fong SJ, Li G, Dey N, Crespo RG, Herrera-Viedma E. Finding an accurate early forecasting model from small dataset: a case of 2019-nCoV novel coronavirus outbreak. Int J Interact Multimed Artif Intell 2020;6(1):132. doi:10.9781/ijimai.2020.02.002.

[11] https://static1.squarespace.com/static/5e6d5df1ff954d5b7b139463/t/

5e77d975e08d1b666d1472f9/1584912792215/ICU_one_pager_COVID_v2.6.pdf%3E.%20Accessed%20April%2014,%202020

[12] De Salazar PM, Niehus R, Taylor A, Buckee C, Lipsitch M. Using predicted imports of 2019-nCoV cases to determine locations that may not be identifying all imported cases. medRxiv, https://doi.org/10.1101/2020.02.04.20020495; 2020.

[13] Read JM, Bridgen JRE, Cummings DAT, Ho A, Jewell CP. Novel coronavirus 2019- nCoV: early estimation of epidemiological parameters and epidemic predictions. medRxiv, https://doi.org/10.1101/2020.01.23.20018549; 2020.

[14] Li X, Wang W, Zhao X, et al. Transmission dynamics and evolutionary history of 2019-nCoV. J Med Virol 2020;92:501–11.

[15] https://Our world in data.org accessed on April 15, 2021

[16] COVID-19 Health advisory platform by Ministry of National Health Services Regulations and Coordination. http://covid.gov.pk/stats/pakistan accessed April 14, 2021.

[17] Raza S, Rasheed MA, Rashid MK. Transmission potential and severity of COVID19 in Pakistan. April 2020. doi:10.20944/PREPRINTS202004.0004.V1.

[18] Raji P1 , Deeba Lakshmi GR, COVID-19-19 pandemic Analysis using Regression, medRxiv, 10.1101/2020.10.08.20208991

[19] Chaudhry R M, Hanif A, Chaudhary M, et al. (May 28, 2020) Coronavirus Disease 2019 (COVID-19-19): Forecast of an Emerging Urgency in Pakistan. Cureus 12(5): e8346. DOI 10.7759/cureus.8346

[20] Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus COVID-19-19. PLoS ONE 15(3): e0231236. https://doi.org/10.1371/ journal.pone.0231236

[21] M. Rubaiyat , Subrato, Prajoy, Priya Podder (2020) Data analytics for novel coronavirus disease" Informatics in Medicine Unlocked, https://doi.org/10.1016/j.imu.2020.100374

[22] Xiongwei Zhang, Hager Saleh , Eman M. G. Younis, Radhya Sahal , and Abdelmgeid A. Ali (2020) Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System, Hindawi Complexity Volume 2020, https://doi.org/10.1155/2020/6688912

[23] Saima Dil, Nyla Dil, Zafar H Maken, "COVID-19-19 Trends and Forecast in the Eastern Mediterranean Region With a Particular Focus on Pakistan, Cureus. 2020 Jun, doi: 10.7759/cureus.8582

[24] Lilleri D, Zavaglio F, Gabanti E, Gerna G, Arbustini E (2020) Analysis of the SARS-CoV-2 epidemic in Italy: The role of local and interventional factors in the control of the epidemic. PLoS ONE 15(11): e0242305. https://doi.org/ 10.1371/journal.pone.0242305

[25] Dominic O'Sullivan , Mubarak Rahamathulla2 and Manohar Pawar (2020) The Impact and Implications of COVID-19-19: An Australian Perspective. The International Journal of Community and Social Development 2(2) 134–151. DOI: 10.1177/2516602620937922

[26] Mohd Saqib (2020) Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression mode. Applied intelligence, Springer. https://doi.org/10.1007/s10489-020-01942-7.

[27] Debanjan Parbat, Monisha Chakraborty (2020) A python based support vector regression model for prediction of COVID-1919 cases in India. Chaos, Solitons and Fractals Nonlinear Science, and Nonequilibrium and Complex Phenomena. https://doi.org/10.1016/j.chaos.2020.109942

[28] https://www.ibm.com/cloud/learn/machine-learning

[29] Daneport, T., & Kalakota, R. - The potential for artificial inteligence in helathcare, Future Healtcare Journal, 6(2), 94-98. (2019)

[30] Alimadadi, A., Aryal, S., Mananadhar, I., Munroe, P.B., Joe. B., & Cheng, X. - Artificial intelligence and machine learning to fight COVID-19. Physiological Genomics, 52(4), 200-202, (2020)

[31] Xu X., Jiang X., Ma C., Du P., Li X., Lv S., Yu L., Ni Q., Chen Y., Su Y., Lang G., Li Y., Zhao H., Xu K., Ruan L., Sheng J., Qiu Y., Wu W., Liang T., Li L. - A Deep learning system to screen novel coronavirus disease 2019 pneumonia. Engineering, ISSN 2095-8099 (2020)

[32] https://www.techopedia.com/definition/33695/labeled-data#

[33] https://labelyourdata.com/articles/unlabeled-data-in-machine-learning/

[34]https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/

[35] https://dataaspirant.com/how-decision-tree-algorithm-works/

[36] https://www.geeksforgeeks.org/instance-based-learning/

[37] https://machinelearningmastery.com/clustering-algorithms-with-python/

[38]https://blog.usejournal.com/association-rule-mining-apriori-algorithm-c517f8d7c54c

[39] https://www.investopedia.com/terms/n/neuralnetwork.asp

[40] https://mlcorner.com/linear-regression-vs-decision-trees/

[41] Fred Gibbs (2013) Installing Python Modules with pip, The programming historian, DOI:10.46430/phen0029

[42]https://www.datacamp.com/community/tutorials/installing-jupyter-notebook?

[43] Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd edition (Springer, New York, NY, 2009).

[44] Jhon Hopkins University (JHU). COVID-19 dashboard by the center of Systems Science and Engineering (CSSE)

[45] Matplotlib. Documentation 2020.

[46] Raihan-Al-Masud M, Mondal MRH. Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. PloS One 2020;15(2): e0228422.

[47] Bharati S, Podder P, Mondal R, Mahmood A, Raihan-Al-Masud M. Comparative performance analysis of different classification algorithm for the purpose of prediction of lung cancer. Intelligent systems design and applications. 2018, vol. 941. Advances in Intelligent Systems and Computing; 2020. p. 447–57.

Appendix A

1. The set of commands use to import data in notebook for defined time period are as follows:

```
\cols = confirmed_df.keys()

\confirmed_df = confirmed_df.loc[:, cols[0]:cols[423]]

\deaths_df = deaths_df.loc[:, cols[0]:cols[423]]

\recoveries_df = recoveries_df.loc[:, cols[0]:cols[423]]
```

2. Set of commands for getting daily COVID-19 cases are shown below:

```
\pk_daily_cases = []

\pk_daily_cases = [pk_cases[i] - pk_cases[i-1] for i in range(1, len(pk_cases)) ]

\pk_daily_deaths = []

\pk_daily_deaths = [pk_deaths[i] - pk_deaths[i-1] for i in range(1, len(pk_deaths)) ]

\pk_daily_recoveries = []

\pk_daily_recoveries = [pk_recoveries[i] - pk_recoveries[i-1] for i in range(1, len(pk_recoveries)) ]

\pk_daily_cases = np.array(pk_cases).reshape(-1, 1)

\pk_daily_deaths = np.array(pk_deaths).reshape(-1, 1)

\pk_daily_recoveries = np.array(pk_recoveries).reshape(-1, 1)
```