Assessing The COVID-19 Trends In Pakistan Using Predictive Machine Learning Techniques: An Empirical Study.

By

HIJAB HASSAN

CIIT/FA19-RSE-018/ISB

MS Thesis

in

Software Engineering

COMSATS University Islamabad, Islamabad - Pakistan

Fall, 2022

COMSATS University Islamabad

Assessing the COVID-19 Trends in Pakistan using Predictive Machine Learning Techniques: An Empirical Study.

A Thesis Presented to

COMSATS University Islamabad

In partial fulfillment
of the requirement for the degree of

MS (Software Engineering)

By
HIJAB HASSAN
CIIT/FA19-RSE-018/ISB

Fall, 2022

Assessing the COVID-19 Trends in Pakistan using Predictive Machine Learning Techniques: An Empirical Study.

A Post Graduate Thesis submitted to the Department of Computer Science as partial fulfilment of the requirement for the award of Degree of MS (Software Engineering).

Name	Registration Number
Hijab Hassan	CIIT/FA19-RSE-018/ISB

Supervisor:

Dr. Muhammad Asim Noor,
Assistant Professor, Department of Computer Science,
COMSATS University Islamabad,
Islamabad, Pakistan

Final Approval

This thesis titled

Assessing the COVID-19 Trends in Pakistan using Predictive Machine Learning Techniques: An Empirical Study.

By

HIJAB HASSAN

CIIT/FA19-RSE-018/ISB

has been approved

For the COMSATS University Islamabad, Islamabad

External Examiner:_____

Dr. XXXX

XXX, XXX Islamabad, Pakistan

Supervisor:_____

Dr. Muhammad Asim Noor

Assistant Professor, Department of Computer Science,

COMSATS University Islamabad, Islamabad

HoD:_____

Dr. Majid Iqbal Khan,

Associate Professor, Department of Computer Science,

COMSATS University Islamabad, Islamabad

Declaration

I HIJAB HASSAN (Registration No. CIIT/FA19-RSE-018/ISB) hereby declare that I have produced the work presented in this thesis, during the scheduled period of study. I also declare that I have not taken any material from any source except referred to wherever due that amount of plagiarism is within acceptable range. If a violation of HEC rules on research has occurred in this thesis, I shall be liable to punishable action under the plagiarism rules of the HEC.

Date: Dec, 2022

HIJAB HASSAN
CIIT/FA19-RSE-018/ISB

Certificate

It is certified that HIJAB HASSAN (Registration No. CIIT/FA19-RSE-018/ISB) has carried out all the work related to this thesis under my supervision at the Department of Computer Science, COMSATS University, Islamabad and the work fulfils the requirement for award of MS degree.

Date: Dec, 2022

Supervisor:

Dr. Muhammad Asim Noor
Assistant Professor, Department of Computer Science

HoD:

Dr. Majid Iqbal Khan
Department of Computer Science

DEDICATION

$\mathcal{D}$edicated

To my beloved parents, siblings and my supervisor Dr Muhammad Asim Noor, who provided me with significant knowledge, and gave me strength when I thought I should give up, by continuously providing me with spiritual and moral support. To my close friends, and office colleagues who encouraged me to complete this research. I also dedicate this research study to my mentor and friend, Junaid Arif Mufti.

And finally, this academic work is whole heartedly dedicated to the young girl who started her academic journey knowing only two words. I am proud of her!

ACKNOWLEDGEMENT

First of all, I would like to thank my supervisor, Dr Muhammad Asim Noor, for his invaluable guidance and feedback. I could not have undertaken this journey without his knowledge and expertise.

This endeavour would not have been possible without the undying support of my parents, Dr Hassan Akhtar and Dr Ghazala Hassan, and siblings. Their love and belief in me have kept my spirits and motivation high during this process. I could not have undertaken this journey without my parents, especially my mother, whose strength has always pushed me to try the impossible and ten years after her own Master's defence, I am about to undertake mine. Thank you for everything.

Additionally, I would like to express my deepest gratitude to my mentor and friend Muhammad Junaid Arif Mufti. He supported and encouraged me to give my best during difficult times. His kind words always lift my spirits and help with my imposter syndrome. And it would be impossible to count all the ways he has helped me in my career and otherwise. Thank you for being there for me, for believing in me and for teaching me so much.

Lastly, a special thanks to my office colleagues especially my work bestie, for their feedback sessions, unlimited food and meme supply, and moral support. Thank you for existing.

ABSTRACT

Assessing the COVID-19 Trends in Pakistan using Predictive Machine Learning Techniques: An Empirical Study

The Coronavirus (also referred to as COVID-19) which started in Wuhan, China on December 2019 has taken the world by storm. Scientists across the globe have used epidemiological models to predict the spread of the virus along with the death rate and make different outbreak predictions. Also, prediction models have been utilized for new policies to control the spread of the virus. Because of the complex and irregular nature of the virus, it has been hard to forecast the trends in different nations specially using conventional mathematical models such as the SIR (Susceptible Infected resistant) model. Therefore, this study analyzes the five waves of COVID-19 that have hit Pakistan since February 2020 using Machine Learning models. Advanced predictive models Predictive models such as Logistic Growth and Autoregressive Integrated Moving Average model (ARIMA) are utilized for predicting and modeling contagion spread trends. The study uses these models to capture the variation in the incidence of daily cases in each province of Pakistan. The time-series data utilized for this study is obtained from the official website of the government of Pakistan; consisting of daily caseload for each region of Pakistan. There are two main contributions of the paper: first, it compares the modeling accuracy of two widely used disease growth models ARIMA, and Logistic Growth Model, in the case of Pakistan. Secondly, it recommends the model better suited for datasets similar to Pakistan, which have fluctuations in numbers. One of the main limitation of this research is that although, one solution for this uncertainty has been the use of Machine Learning predictive techniques, limited data is available for Pakistan. The findings of this research indicated that the logistic model could not model everyday COVID-19 case numbers effectively for the overall pandemic wave, and the model tried to decrease the error, producing an inaccurate plot. However, it showed better results when the waves were divided into smaller sections. The RMSE were less compared to the ARIMA Model. Lastly, the researcher also recommends other models which could be utilized for further modeling of COVID-19 trends in Pakistan.

TABLE OF CONTENTS

Dedication vii

Acknowledgements viii

Abstract ix

List of Figures xiii

List of Tables xv

1 Introduction 1
- 1.1 Overview of Research . . . 2
- 1.2 Motivation . . . 3
- 1.3 Problem Statement . . . 3
- 1.4 Research Objectives . . . 4
- 1.5 Research Questions . . . 4
- 1.6 Research Contribution . . . 4
- 1.7 Research Significance . . . 5
- 1.8 Research Methodology . . . 5
- 1.9 Thesis Structure . . . 6
- 1.10 Chapter Summary . . . 8

2 Preliminary studies 10
- 2.1 ML Techniques used for COVID-19 trends analysis . . . 11
  - 2.1.1 Support Vector Machine . . . 11
  - 2.1.2 Linear Regression model . . . 12
- 2.2 ML Techniques used for COVID-19 trends analysis for this study . 13
  - 2.2.1 The Logistic Growth Model . . . 13
  - 2.2.2 Autoregressive Integrated Moving Average Model (ARIMA) 15
  - 2.2.3 Chapter Summary . . . 19

3 Literature Review 22
- 3.1 Modelling COVID-19 trends using multiple Machine Learning Techniques . . . 25
- 3.2 Predictions and Modelling of Coronavirus cases utilizing the ARIMA Model . . . 35
- 3.3 Predictions and Modelling of Coronavirus cases utilizing the Logistic Growth Model . . . 36
- 3.4 COVID-19 situation in Pakistan . . . 37
- 3.5 The socioeconomic effects of the COVID-19 . . . 44
- 3.6 Outcomes of Literature Review . . . 45
- 3.7 Chapter Summary . . . 45

4 Research Methodology 46

4.1 Research Setting . . . 47
4.2 Research Duration . . . 47
4.3 Methods Used . . . 47
4.4 Sample Size and Technique . . . 47
4.5 Data Analysis . . . 48
4.5.1 Logistic Growth Model . . . 48
4.5.2 ARIMA Model . . . 50
4.6 Dataset(s) . . . 51
4.7 Evaluation method(s) and criteria . . . 51
4.7.1 Model Evaluation . . . 51
4.7.2 Description . . . 52
4.7.3 Modeling . . . 53
4.7.4 Predicting . . . 54
4.7.5 Control . . . 54
4.7.6 Chapter Summary . . . 54

5 Proposed Solution 55
5.1 Proposed Solution . . . 56
5.1.1 Secular trend movement . . . 56
5.1.2 Seasonal movement . . . 56
5.1.3 Cyclic movement . . . 56
5.1.4 Unpredictable movement . . . 58
5.1.5 Stationary or fixed time series . . . 58
5.1.6 Achieving Stationarity . . . 58
5.1.6.0.1 Regular Differencing . . . 58
5.1.6.0.2 Seasonal Differencing . . . 58
5.1.7 Autocorrelation and partial autocorrelation function . . . 59
5.1.7.0.1 ACF (Autocorrelation Function) . . . 59
5.1.7.0.2 The sample data distribution of coefficients of the Autocorrelation Function . . . 59
5.1.7.0.3 PACF (Partial Autocorrelation Function) 59
5.1.7.0.4 SPACF (Sample Partial Autocorrelation Function) . . . 59
5.1.8 Models of time series . . . 60
5.1.8.0.1 The AR(p) (Autoregressive model) . . . 60
5.1.8.0.2 The MA(q) (Moving Average model) . . 60
5.1.8.0.3 ARMA (Autoregressive Moving Average) model . . . 60
5.2 ARIMA (Autoregressive Integrated Moving Average) model . . . 61
5.2.0.0.1 Seasonal Autoregressive Integrated Moving Average model . . . 61
5.3 The Box-Jenkins Method . . . 61
5.3.1 Identification of the tentative . . . 62
5.3.2 Detection of Stationarity . . . 62
5.3.3 Evaluating Stationarity of the time-series . . . 63
5.3.3.0.1 ADF (Augmented Dickey Fuller) . . . 63
5.3.3.0.2 Differencing to achieve Stationarity . . . 64

5.3.4 The selection criteria for choosing a model . . . 64
5.3.4.0.1 AIC (Akaike Information Criteria) . . . 65
5.3.4.0.2 BIC (Schwarz Bayesian Information Criteria) . . . 65
5.3.4.0.3 R-Square Criteria . . . 66
5.4 Estimating the parameters of the tentative model . . . 66
5.5 Model diagnostics . . . 66
5.5.1 Autocorrelation Function and Partial Autocorrelation Function plots of residuals . . . 67
5.5.2 Normality Test . . . 67
5.5.3 Ljung Box Chi-Square Test . . . 67
5.5.4 Jarque-Bera Test . . . 67
5.6 Model Forecast . . . 69
5.7 Measuring Forecasting Accuracy . . . 69
5.7.1 MAPE (Mean Absolute Percentage Error) . . . 69
5.7.2 RMSE (Root Mean Square Error) . . . 69
5.7.3 MAE (Mean Absolute Error) . . . 71
5.8 Logistic Growth Model . . . 71
5.9 Chapter Summary . . . 72

6 Experimental Results and Evaluation 73
6.1 Dataset . . . 74
6.2 Experimental Settings . . . 75
6.3 Results . . . 79
6.3.1 ARIMA Model Fit for all over Pakistan . . . 79
6.3.2 ARIMA Error Analysis . . . 84
6.3.3 Logistic Growth Fit . . . 85
6.3.4 Logistic Error Analysis . . . 88
6.3.5 Chapter Summary . . . 89

7 Conclusion and Future Work 90
7.1 Conclusion . . . 91
7.2 Future work . . . 91

8 References 92

Appendices 95
.1 Appendices . . . 96

LIST OF FIGURES

1.1 Systematic Literature Review Workflow . . . 6
1.2 Methodology Workflow . . . 7
1.3 Structure of Dissertation . . . 8

2.1 Support Vector Machine Hyper planes . . . 11
2.2 The flowchart of Support Vector Machine based classification . . . 12
2.3 The Linear regression model . . . 13
2.4 The use of the Logistic Growth model. . . . 14
2.5 The Logistic growth modeling of Coronavirus and Severe Acute Respiratory Syndrome in China [36] . . . 15
2.6 Plotting the total confirmed cases of Coronavirus in China [39] . . . 16
2.7 Plotting the total confirmed cases of Coronavirus in China in 2020 [39] . . . 16
2.8 The total number of confirmed cases versus the total number of casualties In China in 2020 [39]] . . . 17
2.9 The Q-Q plot of confirmed Coronavirus cases in China [39] . . . 17
2.10 The Q-Q plot of deaths Coronavirus in China [39] . . . 18
2.11 White noise time-series of confirmed Coronavirus cases in China. . . . 19
2.12 White noise time-series of fatal cases . . . 19
2.13 White noise time-series confirmed Coronavirus case versus confirmed cases . . . 20
2.14 White noise time-series fatal case versus fatal cases . . . 20
2.15 Seasonal decomposition of confirmed Coronavirus cases in 2020 in China . . . 21

3.1 Possible solution approaches for predicting COVID-19 trends across various countries . . . 23
3.2 Systematic Literature Review for the research study . . . 24
3.3 Accuracy graph of the proposed model . . . 43
3.4 Possible solution approaches for predicting COVID-19 trends across various countries . . . 44
3.5 Cases reported in Pakistan . . . 44

4.1 Workflow for methods utilized in study . . . 48
4.2 Sampling Strategy . . . 48
4.3 LGM Algorithm . . . 49
4.4 ARIMA Algorithm . . . 52
4.5 ARIMA Algorithm (Continued) . . . 53
4.6 Data Pre-Processing . . . 53

5.1 Proposed Solution . . . 57
5.2 Seasonal Autoregressive Integrated Moving Average model . . . 62
5.3 The Box-Jenkins Methodology . . . 63
5.4 The Augmented Dickey Fuller test . . . 65
5.5 The Ljung box test . . . 68

5.6 The Mean Absolute Percentage Error 70
5.7 The Root Mean Square Error 70
5.8 The Mean Absolute Error 71

6.1 Visualization of daily COVID-19 cases in Pakistan 75
6.2 Visualization of daily COVID-19 cases in Punjab 75
6.3 Visualization of daily COVID-19 cases in Sindh 76
6.4 Visualization of daily COVID-19 cases in Khyber Pakhtunkhuwa . . 76
6.5 Visualization of daily COVID-19 cases in Baluchistan 77
6.6 Visualization of daily COVID-19 cases in Gilgit Baltistan. 77
6.7 Visualization of daily COVID-19 cases in Azad Jammu and Kashmir 78
6.8 Visualization of daily COVID-19 cases in Islamabad. 78
6.9 ARIMA FIT for 874 days of data for daily COVID-19 cases in Pakistan 80
6.10 ARIMA FIT for 874 days of data for daily COVID-19 cases in Punjab 80
6.11 ARIMA FIT for 874 days of data for daily COVID-19 cases in Sindh 81
6.12 ARIMA FIT for 874 days of data for daily COVID-19 cases in Khyber Pakhtunkhuwa 81
6.13 ARIMA FIT for 874 days of data for daily COVID-19 cases in Baluchistan 82
6.14 ARIMA FIT for 874 days of data for daily COVID-19 cases in Gilgit Baltistan. 82
6.15 ARIMA FIT for 874 days of data for daily COVID-19 cases in Azad Jammu and Kashmir 83
6.16 ARIMA FIT for 874 days of data for daily COVID-19 cases in Islamabad 83
6.17 Logistic Growth Modeling for 1st Wave 85
6.18 Logistic Growth Modeling for 2nd Wave. 86
6.19 Logistic Growth Modeling for 3rd Wave. 86
6.20 Logistic Growth Modeling for 4th Wave. 87
6.21 Logistic Growth Modeling for 5th Wave. 87

1 LGM Algorithm 97
2 ARIMA Algorithm 98
3 ARIMA Algorithm (Continued) 99
4 ARIMA Algorithm (Continued) 100
5 Data Pre-Processing 101

LIST OF TABLES

3.1 Review Matrix for Literature Review 25

5.1 Hypothetical Scheme for ACF and PACF for non-seasonal time-series data 64
5.2 Hypothetical Scheme for ACF and PACF for seasonal time-series data 64

6.1 COVID-19 Waves mapped out on the dates that occurred. 74
6.2 Summary of ARIMA Models used on time series data. 79
6.3 Error Analysis for ARIMA Model for each wave. 84
6.4 Error Analysis for ARIMA Model for overall wave for each region. . 84
6.5 Error Analysis for Logistic Growth Model for each wave. 88

Chapter 1

Introduction

1.1 Overview of Research

The Coronavirus (also known as COVID-19) started which started in Wuhan, China on December 2019 spread like fire across the world leading to possibly the worst public health crisis since the spread of Influenza in 1918. The pandemic has over 118 million confirm cases along with 2.63 million deaths and 67.1 million recovered survivors till March 2020 [1]. In Pakistan alone, there have been 13,377 deaths and 566,493 recovered cases [2]. The first COVID-19 case in Pakistan was reported in Karachi, Sindh and later confirmed by the Ministry of Health. The COVID-19 pandemic has been aggressive and relentless in its spread. Scientists around the globe have utilized epidemiological models to predict the spread of the virus along with the mortality rate and make other outbreak predictions. Moreover, prediction models have been used for new policies to control the spread of the virus and evaluate already implemented strategies such as the lockdown until a viable treatment can be created and administered to the general public. However, this has been difficult in the case of COVID-19. Due to the complex and irregular nature of the virus, it has been difficult to predict the trends in various countries, especially with traditional models such as the SIR (Susceptible Infected Resistant) model. These models have been unable to provide the higher performance required in certain regions of the globe for effective policy-making decisions [3]. As a result, there has been a global need to provide effective analytical solutions to deal with the drawback of SIR models. One such solution is the use of Machine Learning predictive techniques. Due to the uncertainty and complexity surrounding the COVID-19 spread, multiple Machine Learning techniques have found great success in various countries. Some of these include multiple linear regression, ARIMA (Autoregressive Integrated Moving Average), forecasting using available data, support vector machine, and other predictive models as well [3]. The common purpose of these models is to provide support to the administration in terms of providing plans for peak time, the spreading scale, and resource allocation planning. This would help prevent and control the spread of the virus. Therefore, this study will help assess which predictive model will be best suited for the COVID-19 trends in Pakistan. This will play a significant role in understanding which policies and strategies that were employed by the Pakistani government were most suitable to curb the spread of the virus. Moreover, the prediction techniques would be significant in helping the administration deal with the third wave of the virus currently spreading through the country.

1.2 Motivation

Several researchers worldwide made an honest effort to foresee or estimate the evolution of Coronavirus in their nations utilizing several mathematical, statistical, computing, and deep learning models and methods. In order to understand and more specifically, to predict the spread of Coronavirus, a few explorations have been conducted because of the presence of this new virus, utilizing different mathematical models. These mathematical models, which depend on differential equations, have the significant disadvantage of neglecting to create practical and helpful outcomes, especially for complicated disease frameworks, such as Coronavirus. On the contrary, time series forecasting outbreaks like Coronavirus utilizing mathematical models as well as deep learning are of crucial significance, particularly in these phenomenal times. It does not just provide a long or short-term overview of the Coronavirus situation in a specific nation, such as Pakistan. However, also helps decision-makers to carefully make the right decisions or moves.

1.3 Problem Statement

Due to the socio-economic impact of the COVID-19 pandemic, it has become necessary to gauge the progress of the virus around the globe [4]. Traditional prediction models such as the SIR models demonstrate low accuracy when predicting COVID-19 trends due to its complex nature [5] [6]. As a result, there has been a global need to provide effective analytical solutions to deal with the drawback of SIR models [7]. Machine Learning predictive models such as Linear Regression [8], Logistic Growth Model [9], ARIMA [10], Decision Trees [11] and Random Forest [13] have shown to be effective solutions in predicting COVID- 19 trends in various countries such as Iceland, Netherlands, Brazil, USA and India. These have been successful in allowing governments to deal with major policy making related to the pandemic [14] [15] [16] [17]. In Pakistan, the spread of the COVID-19 has been curbed by effective policy making decisions taken by the Pakistani government [18] [19]. However, there is little evidence about which machine learning models [20] are highly effective in modelling infectious disease patterns with high variability in daily caseloads in Pakistan along with which policy has been the most effective in controlling the pandemic.

1.4 Research Objectives

This section should explain the objective of the research The primary purpose of this study is to compare and contrast the various predictive machine learning models available for the COVID-19 trends assessment for Pakistan. The study will analyze which model is best suited to the Pakistani data for COVID-19 and give useful insight into the strategies used by the government to control the spread of infection. In addition to this, an analysis will be carried out with other countries to find out why the COVID-19 cumulative cases and deaths were less compared to other countries. The following objectives are to be achieved for this study:

- To determine which Machine Learning model better models COVID-19 trends in Pakistan.
- To determine which model produces a higher accuracy for the variable trends of Pakistan.

1.5 Research Questions

The following research questions are to be addressed in this study:

- RQ-1: Which predictive Machine Learning model is most suitable for modelling COVID-19 trends in Pakistan?
- RQ-2: Which predictive Machine Learning model is most suitable for predicting new COVID-19 trends in Pakistan?
- RQ-3: Which predictive Machine Learning model gives a higher accuracy for the COVID-19 trends in Pakistan?

1.6 Research Contribution

Research contribution is:

- The study will analyze which model is best suited to the Pakistani data for COVID-19.
- The proposed method involves two steps. In the first step, training and modeling two growth models to empirically study high variability data trends in Pakistan. The second step is to carry out an error analysis for the models to compare accuracy.

- As a result, error analysis is used to average caseload data and results are discussed with recommendations for future studies.

1.7 Research Significance

The research is significant because it will guide healthcare professionals and the government of Pakistan regarding how they monitor and control the number of positive Coronavirus cases in Pakistan in the future, which will eventually help to decrease the number of deaths caused by Coronavirus. Moreover, it will allow others to model caseload trends for infectious diseases with highly variable case trends.

1.8 Research Methodology

The literature review has demonstrated that the time series machine learning models are the most effective in analyzing COVID-19 trends. Moreover, for a dataset like Pakistan's which is irregular and has a non-linear trend. The models include ARIMA and Logistic Growth model.

The research study will carry out a systematic literature review to find out which prediction models are already used in the previous studies. The models will be short-listed and then trained for the required data set. The research study will consider two main Machine Learning models for the data set collected for Pakistan. Each technique will be modelled according to the dataset collected and the accuracy of the predicted values will be compared with the actual value. The accuracy of the models will be analyzed using the mean square error and mean absolute error [2]. Figure 1.1 summarizes the research study's systematic literature review. Prediction plots will be used to assess the accuracy of the models as well and determine which model best fits the COVID-19 data for the study. Each model will have a pre-defined.

The data collected from the models will be plotted for a better graphical representation. Figure 1.2 shows the data analysis process for this study. A descriptive data analysis will be carried out to show the general trends in the data. Later, scatter plots will be discussed to determine the best model. Accuracy for each model will be determined using the mean square error and mean absolute error as they were considered to be most commonly and effectively used by previous studies.

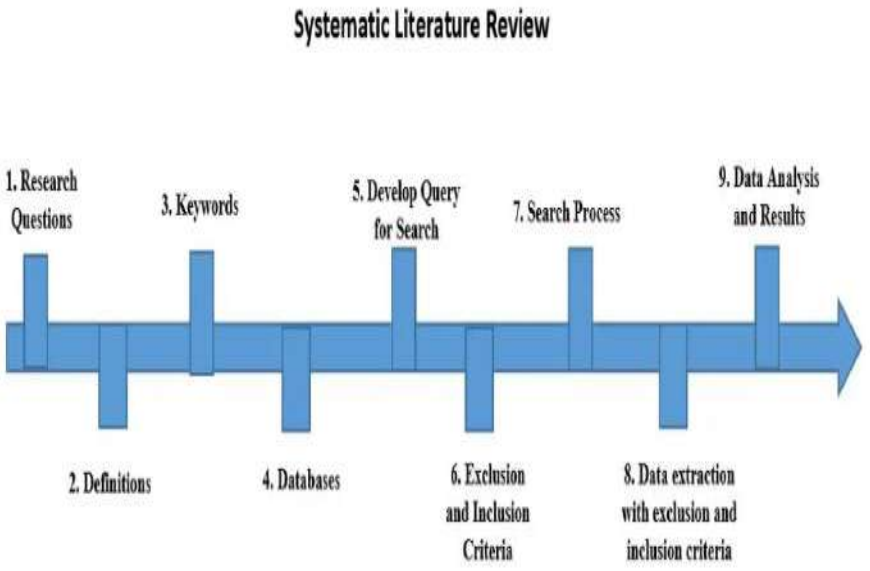


FIGURE 1.1: Systematic Literature Review Workflow

The data collected from the models will be plotted for a better graphical representation. A descriptive data analysis will be carried out to show the general trends in the data. Later, scatter plots will be discussed to determine the best model. Accuracy for each model will be determined using the mean square error and mean absolute error as they were considered to be most commonly and effectively used by previous studies.

1.9 Thesis Structure

The remaining sections of this research paper are ordered in the following style: Machine Learning models have been presented in Chapter 2. Chapter 3 consists of the analysis of preliminary studies. Chapter 4 dives deep into the systematic literature review carried out for this study. Conceptual framework, results, and discussions are included in Chapter 5 and Chapter 6. Lastly, in Chapter 7, we provided Conclusions and Future works. The structure of this dissertation is provided in Figure 1.3.

Chapter 1: Introduction-this chapter presents the research study, and details about the significance of the study and its applicability are discussed. Moreover, the research objectives, research questions, and research contributions are also discussed

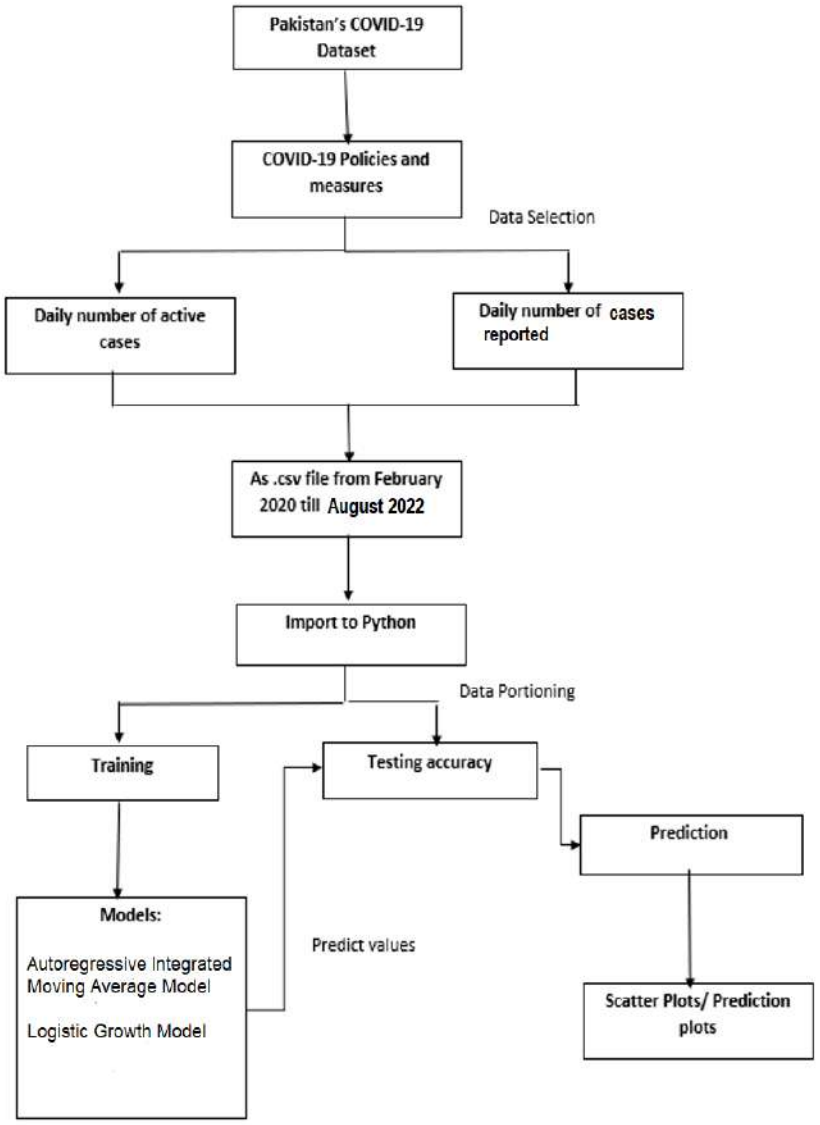


FIGURE 1.2: Methodology Workflow

in detail.

Chapter 2: Preliminary Studies- this chapter presents details about machine learning algorithms that were assessed for this research for the prediction of endeavor.

Chapter 3: systematic literature review- this chapter discusses past studies and recent literature studies in detail to find the possible research gaps in the exploration.

Chapter 4: Methodology- this chapter discusses the methodology used to address the research questions.

Chapter 5- Conceptual Framework- this chapter describes the proposed conceptual framework and explains every segment of this framework thoroughly.

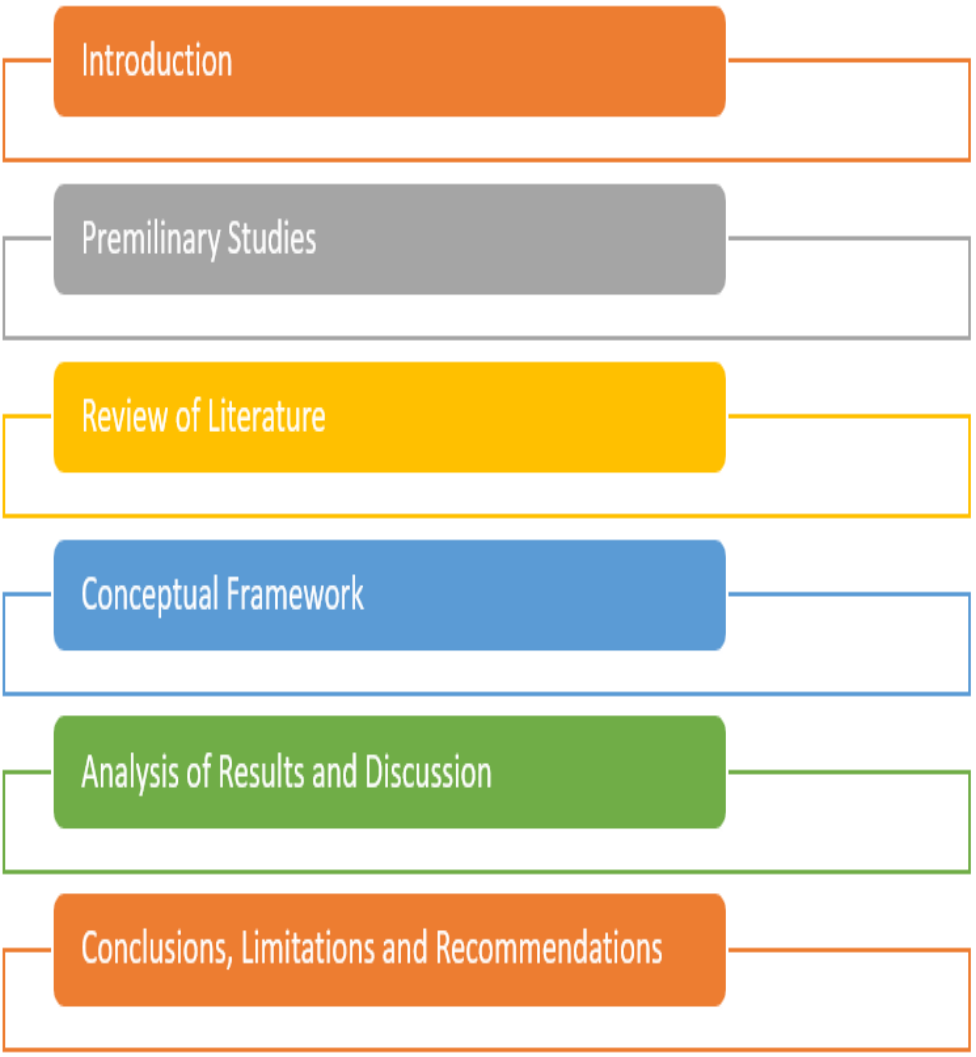


FIGURE 1.3: Structure of Dissertation

Chapter 6: Results and Discussions- this chapter provides an in-depth interpretation of the findings. In addition to this, different validation measures are also discussed.

Chapter 7: Conclusions and Future Work- this is the final chapter of this research paper, which provides the conclusion and guides the future view of the researchers in the area of predictive ML (Machine Learning) models.

1.10 Chapter Summary

This chapter depicts a thorough introduction to the proposed research work. This study proposed using two different predictive machine learning models, such as

the Logistic Growth Model and the AIMA (Autoregressive Integrated Moving Average) model to assess the trends of Coronavirus in Pakistan. The machine learning models utilized in past studies were not able to produce accurate results when assessing the trends of COVID-19. In this study, Autoregressive Integrated Moving Average and Logistic Growth Model are proposed to capture the variation in the incidence of daily cases in every province of Pakistan. The first section of the Introduction chapter provides a comprehensive overview of the study. After that, the motivation behind conducting this study is discussed. In the subsequent sections, the problem statement is depicted and the research questions are raised.

Chapter 2

Preliminary Studies

This chapter is based on the preliminary studies to give a better understanding of the model related to this research. This section provides an examination of the basic concepts of the possible Machine Learning techniques that might have been utilized in this research, such as Support Vector Machine and techniques that were already used in this research. For example, the Logistic Growth model and Autoregressive Integrated Moving Average model. This chapter gives a better understating of the main concepts of the model and the techniques related to this research work.

2.1 ML Techniques used for COVID-19 trends analysis

2.1.1 Support Vector Machine

One of the most commonly utilized modern ML techniques is SVM (Support Vector Machine). In Machine Learning, it is a type of supervised learning model with related learning algorithms that analyze data to conduct regression and classification analysis. Also, to perform linear classification, Support Vector Machines could

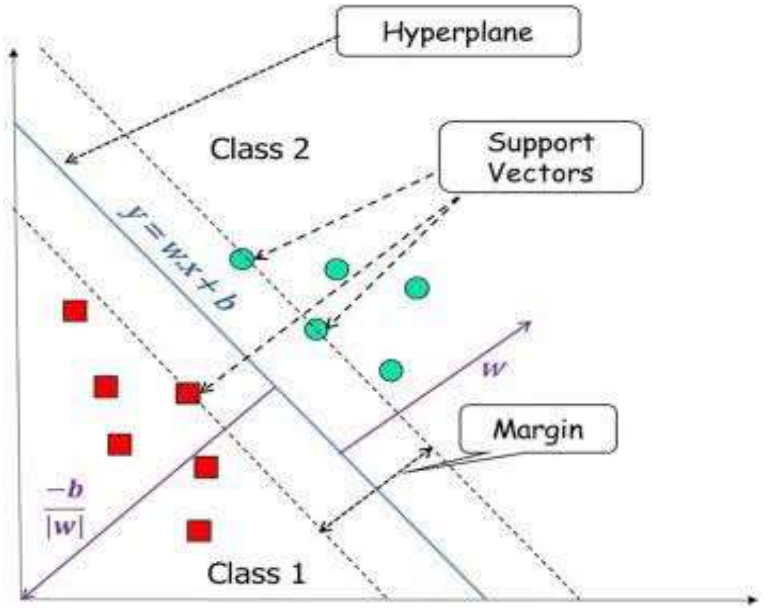


FIGURE 2.1: Support Vector Machine Hyper planes

effectively conduct a non-linear classification analysis utilizing what is known as "Kernel Trick", to map their inputs into high-dimensional feature areas. It fundamentally draw margins between different classes as shown in Figure 2.2. The

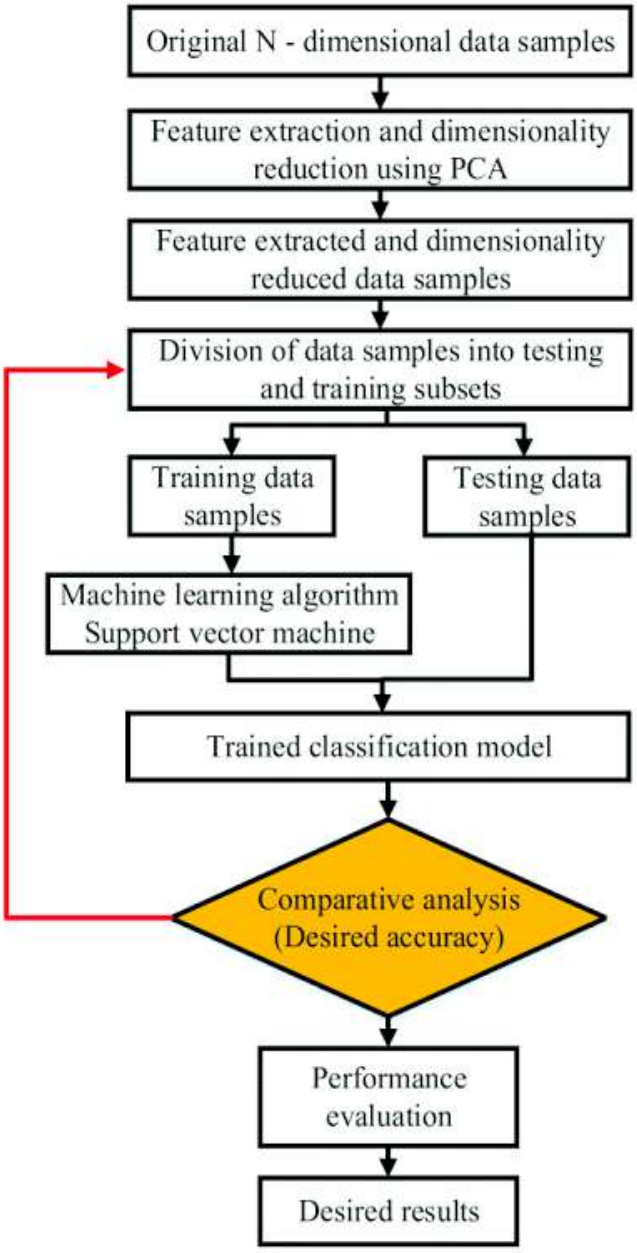


FIGURE 2.2: The flowchart of Support Vector Machine based classification

margins are drawn in such a manner that there is maximum distance between the classes and the margin, and thus, reducing the classification error.

2.1.2 Linear Regression model

Linear regression analysis is utilized to forecast a factor's value that depends on the value of another factor. The factor one needs to predict is referred to as the dependent factor. The factor one has utilized to predict the value of another factor is known as the independent factor. This type of model is used to estimate the

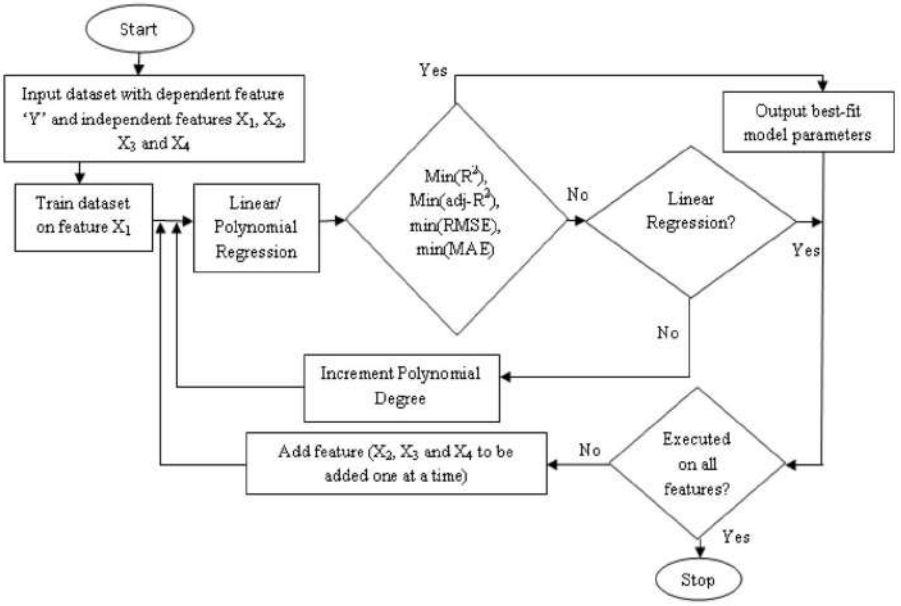


FIGURE 2.3: The Linear regression model

coefficient value of a linear equation, including at least one independent factor that best predicts the linear equation, including at least one independent factor that best predicts the dependent factor's value. Linear regression fits the straight line that minimizes the differences between forecasted and substantial result values as shown in Figure 2.3 [26].

2.2 ML Techniques used for COVID-19 trends analysis for this study

2.2.1 The Logistic Growth Model

The Logistic Model is derived from the modeling of the populace growth in an environment [35]. In order to make improvements to the Malthus population model, Pierre François Verhulst created a logistic formula:

$$\frac{dQ}{dt} = rQ(\frac{1-Q}{K}) \tag{2.1}$$

where K, Q, and r signify the size of the population, the intrinsic growth rate as well as the maximum size of the populace that the surroundings can carry, individually. dQ / dt denotes the population growth. K and r are constant numbers and the value of Q is derived with time to create a curve that is S-shaped within this Logistic formula as shown in Figure 2.4 Presently, logical functions have been applied inside a time series forecast issue past their environmental roots, similar

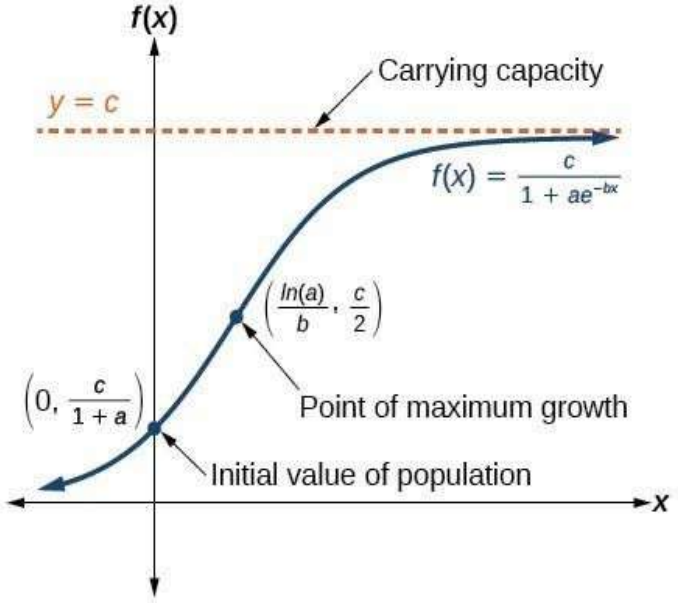
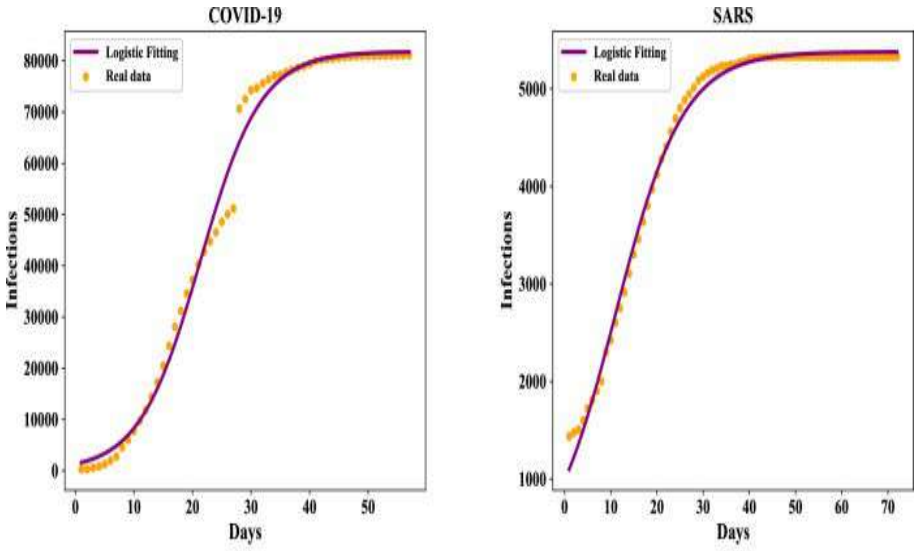


FIGURE 2.4: The use of the Logistic Growth model.

to epidemiology modeling. Logistic growth is described by the rising growth rate during the starting period. However, declining growth in the next phase, as one gets near the maximum size of the population. [36] utilized the Richards model (a kind of Logistics model) to fit the total number of Severe Acute Respiratory Syndrome reported every day in Beijing, Singapore, and Hong Kong. As it is illustrated in Figure 2.5, the total number of Severe Acute Respiratory Syndrome and Coronavirus in China is usually a curve that is S-shaped, which can be depicted by Logistic fitting. Toward the start of the COVID-19 pandemic, individuals did not adopt rigorous measures, as well as the initial number of individuals infected by Coronavirus was not high, so the number of infections gradually increased.

Once the base of the infection surpassed a specific ratio, the pandemic situation indicated a substantial exponential growth trend, and afterwards, with the help of government regulations and the cooperation from the general public, the situation of the pandemic slowly decreased the pace at which the virus spread, eventually attained the minimal number of total infected individuals. The main point is the time at which the curve of the cumulative situation turned, for example, when the rapid increase in the number of infected cases was replaced by a slow increase. Since the point of infection indicates the point at which the everyday number of infected cases increased and reached its maximum value, this point denotes a basic defining moment at which transmission of the virus starts to decrease. However, the data included this critical point as well as the time span shortly subsequently, the fitting of the curve and prediction of the future number of infected cases would



(a) COVID-19 (b) SARS

FIGURE 2.5: The Logistic growth modeling of Coronavirus and Severe Acute Respiratory Syndrome in China [36]

be relatively precise as shown in Figures 2.6-2.13

2.2.2 Autoregressive Integrated Moving Average Model (ARIMA)

Autoregressive Integrated Moving Average model is a popular and flexible category of prediction models that utilizes recorded data in order to make predictions. This model is a crucial prediction method that could act as a beginning point for logically complicated models [39]. It works efficiently when the data shows consistent examples that are predictable after a while with a base estimation of abnormalities. The Autoregressive Integrated Moving Average approach tries to depict improvements in a fixed time series as a component of what is assigned as "Autoregressive and Moving Normal" boundaries. These are referred to as Autoregressive boundaries and MA (Moving Average) boundaries. We acknowledge that time is a discrete factor, Z_t indicates the observations at t time and lastly, t indicates the zero-mean random noise term at t. The Moving Average mode utilizes this process:

$$Z_t = X\gamma_i\epsilon_{t-1} + \epsilon_t \tag{2.2}$$

where γ_i represents the coefficient, identical to Moving Average (n) models, Autoregression model, reported by AR(m):

$$Z_t = X\delta_i Z_{t-1} + \epsilon_t \tag{2.3}$$

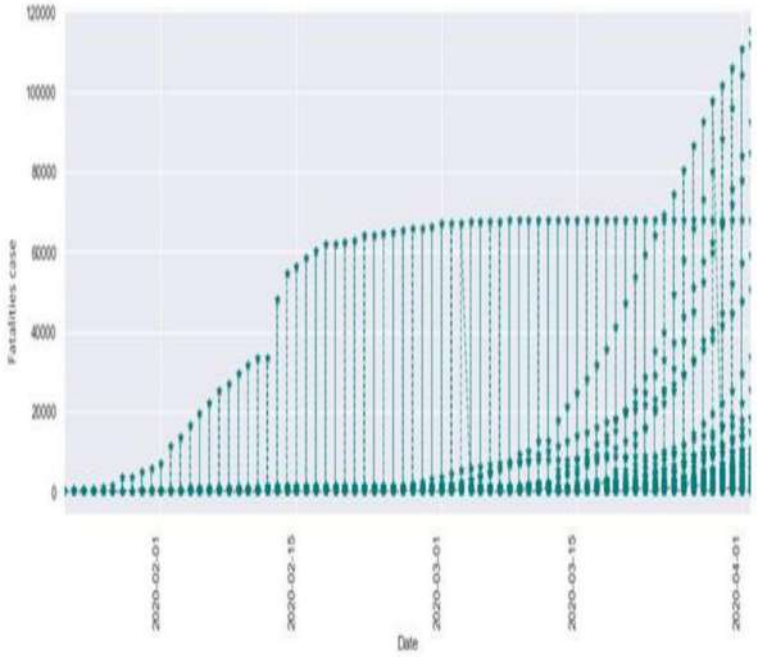


FIGURE 2.6: Plotting the total confirmed cases of Coronavirus in China [39]

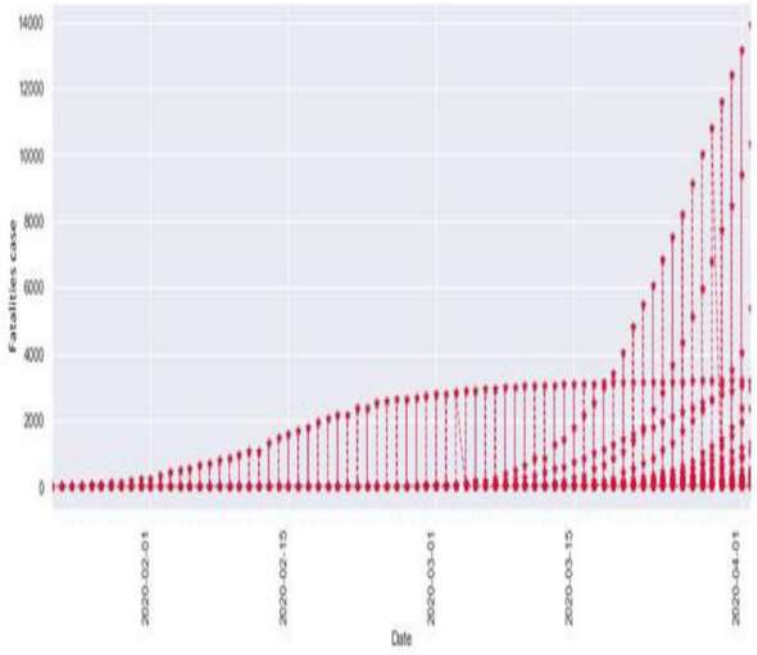


FIGURE 2.7: Plotting the total confirmed cases of Coronavirus in China in 2020 [39]

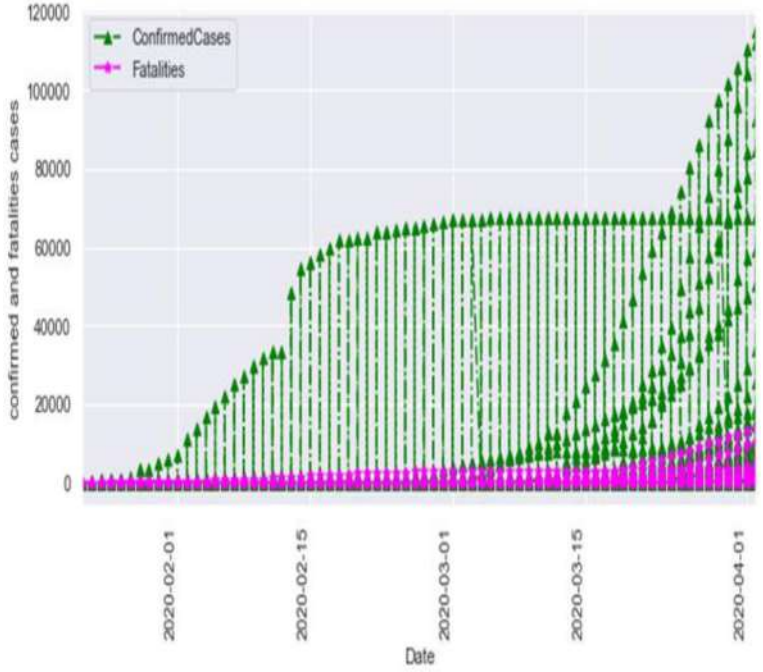


FIGURE 2.8: The total number of confirmed cases versus the total number of casualties In China in 2020 [39]]

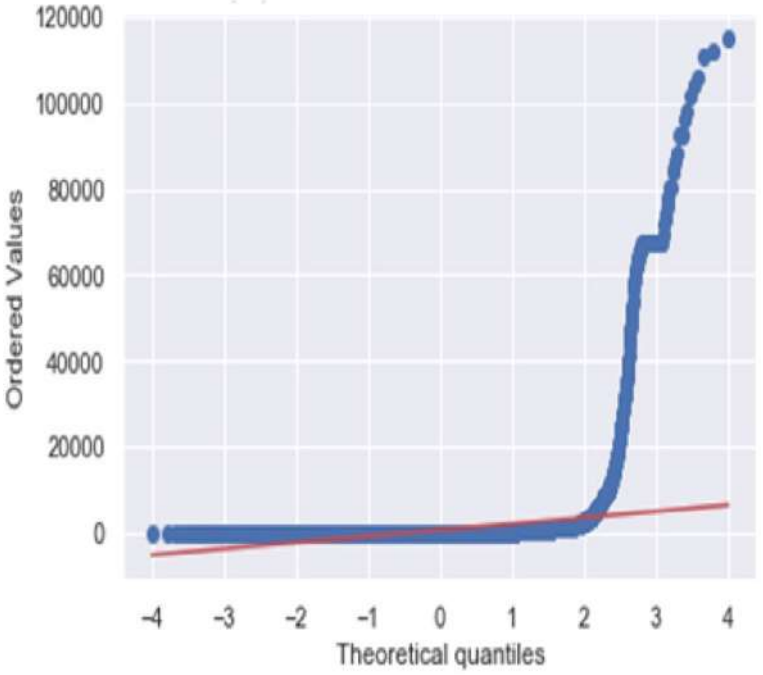


FIGURE 2.9: The Q-Q plot of confirmed Coronavirus cases in China [39]

Z_t is a noisy linear combination of the m observations that were taken previously. An increasingly developed model is the Autoregressive Integrated Moving Average (m, n), a combination of Autoregressive (m), and Moving Average (n) with a

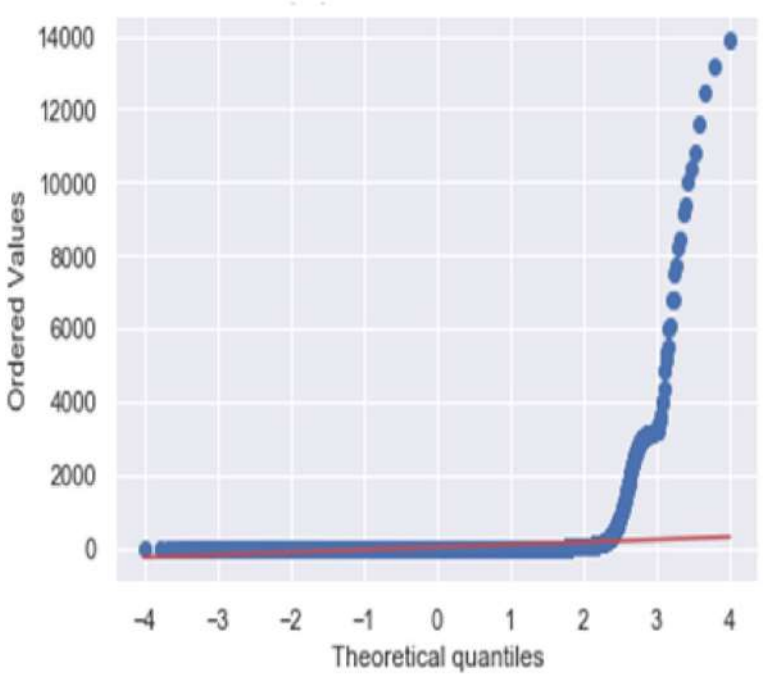


FIGURE 2.10: The Q-Q plot of deaths Coronavirus in China [39]

decreased framework and provides a fixable demonstrating structure. This model predicts that Z_t is created using the following equation:

Where t is the zero-mean noise term. In case, we are including an imperative to the Autoregressive (m) component, it guarantees a stationary procedure. A stationary and non-vertible Autoregressive Integrated Moving Average (m, n) model might be described either as an infinite Autoregressive model (AR(∞)) or an infinite Moving Average model (MA(∞)).

For the Autoregressive Integrated Moving Average model, we could calculate the first-order difference of Z_t by $\nabla Z_t = Z_t Z_{t1}$ and the second-order difference of Z_t by $\nabla 2 Z_t = \nabla Z_t \nabla Z_{t1}$ in such a way that the series of ∇Z_t fulfils an Autoregressive Integrated Moving Average (m, n). One could see that the series of Z_t fulfils the Autoregressive Integrated Moving Average (m, d, n):

For the Autoregressive Integrated Moving Average model, we could calculate the first-order difference of Z_t by $\nabla Z_t = Z_t Z_{t1}$ and the second-order difference of Z_t by $\nabla 2 Z_t = \nabla Z_t \nabla Z_{t1}$ in such a way that the series of ∇Z_t fulfils an Autoregressive Integrated Moving Average (m, n). One could see that the series of Z_t fulfils the Autoregressive Integrated Moving Average (m, d, n): which are determined by three order parameters terms m, d, n with particular weights vector δ m and γ. Making predictions using an Autoregressive Integrated Moving Average (m, d, n) is an inversion of the differential equations. Presuming the time-series sequence

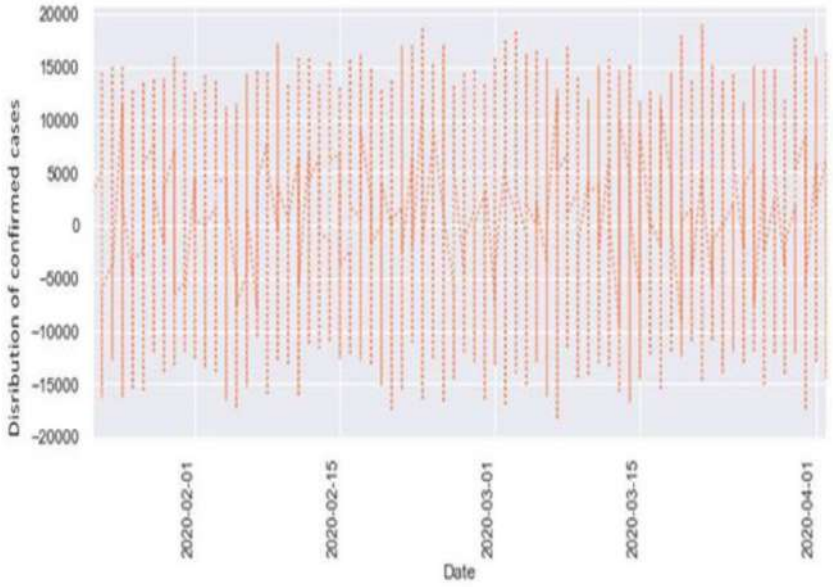


FIGURE 2.11: White noise time-series of confirmed Coronavirus cases in China.

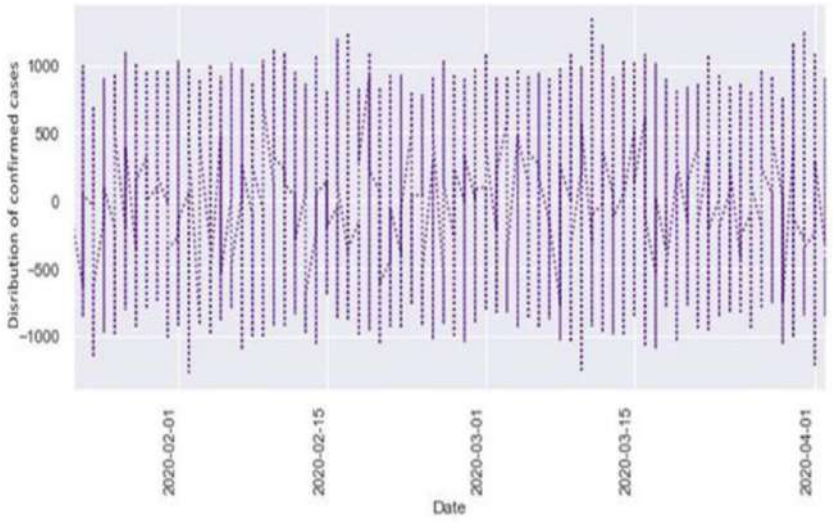


FIGURE 2.12: White noise time-series of fatal cases

Z_t fulfils Autoregressive Integrated Moving Average (m, d, n), we can anticipate the dth order differential of observations at time $t+1$ as and afterwards forecast the observations at time $t+1$ as $Z_t^{\sim}$:

$$Z_t^{\sim} = \nabla Z_t^{\sim} + \Sigma \nabla^i Z_{t1} \tag{2.4}$$

2.2.3 Chapter Summary

This chapter has described the different Machine Learning model that could have used in this research as well as the models that were actually used in this research.

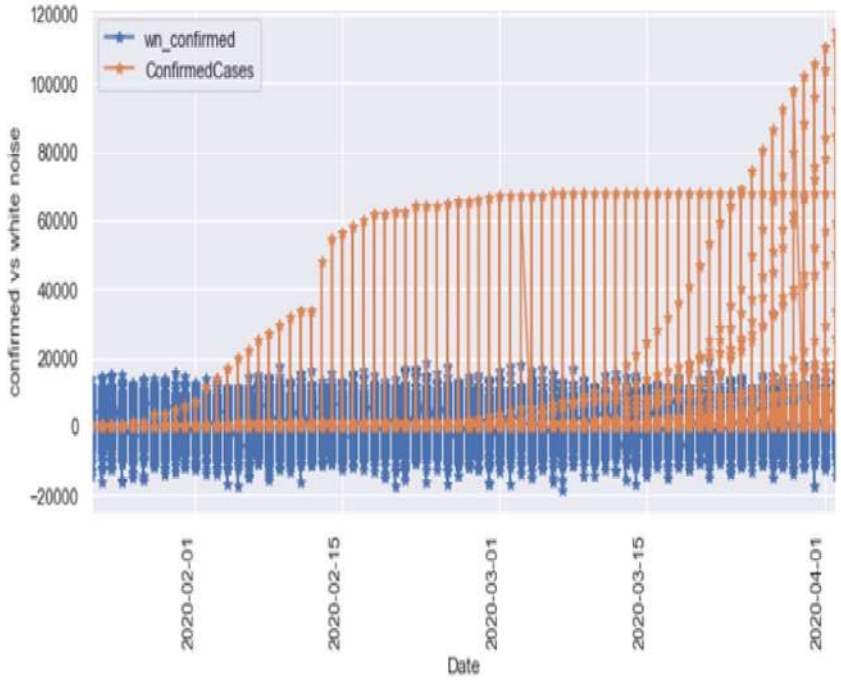


FIGURE 2.13: White noise time-series confirmed Coronavirus case versus confirmed cases

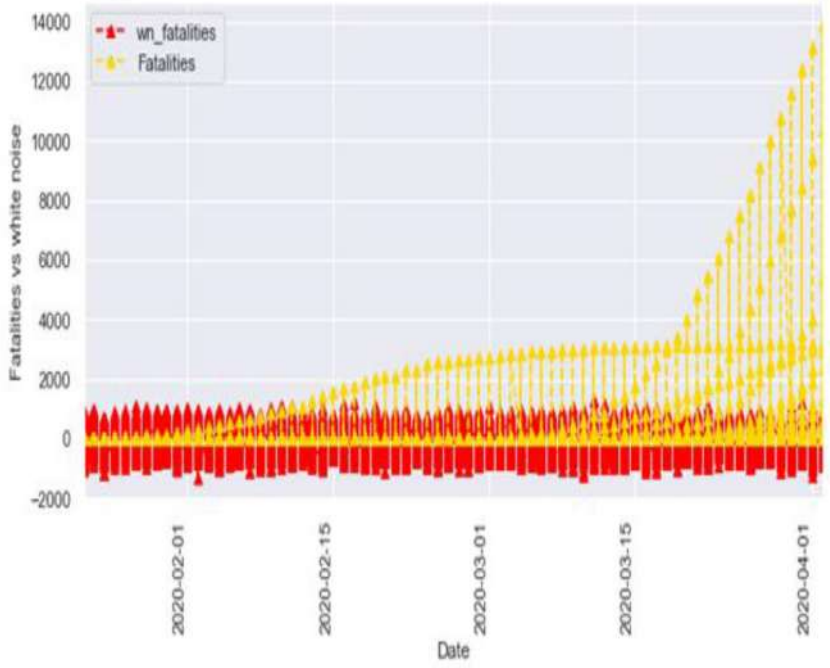


FIGURE 2.14: White noise time-series fatal case versus fatal cases

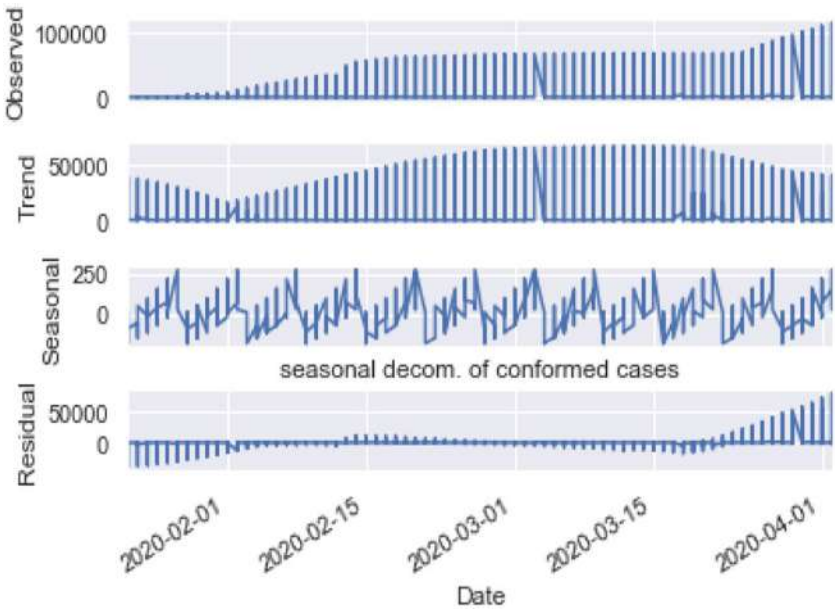


FIGURE 2.15: Seasonal decomposition of confirmed Coronavirus cases in 2020 in China

We provided details about the Logistic Growth model, the Autoregressive Integrated Moving Average and the Support Vector Machine. In the next chapter, we will present relevant studies related to the research topic.

Chapter 3

Literature Review

In this research work, there are three dimensions of the literature review. The first one analysis of machine learning tools for COVID-19 trends. The second one is modelling of COVID-19 trends using ARIMA Model and Logistic Growth Model. The third is the COVID-19 situation in Pakistan.

The COVID-19 phenomena have shaken the world. Many countries around the world have tried implementing policies that may subdue the progress of the virus. All over the world, researchers have been working on predicting and forecasting the number of future cases, deaths and recoveries in order to foresee the pandemic situation to make effective future policies [3]. Multiple machine learning models have been used to gauge the intensity of the COVID-19 pandemic. Most of the literature assessed for this study has been displayed in the form a review matrix in Table 3.1. Figure 3.1 shows the percentage contribution of the Machine Learning solutions for the COVID-19 pandemic forecasting problem.

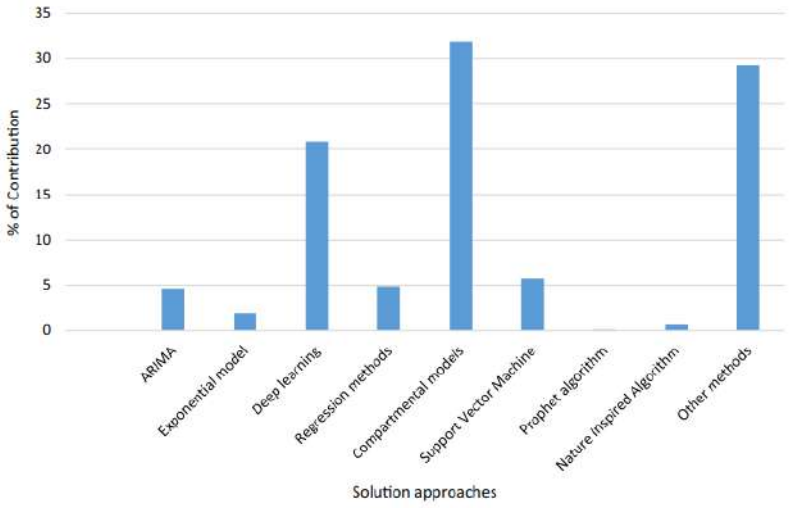


FIGURE 3.1: Possible solution approaches for predicting COVID-19 trends across various countries

The literature review showed that Support Vector Machine, Logistic and Linear Regression, ARIMA, Bayesian Regressor model etc. were used for forecasting purposes (Daniyal, 2020). Figure 3.2 show the Systematic Literature Review carried out for this study.

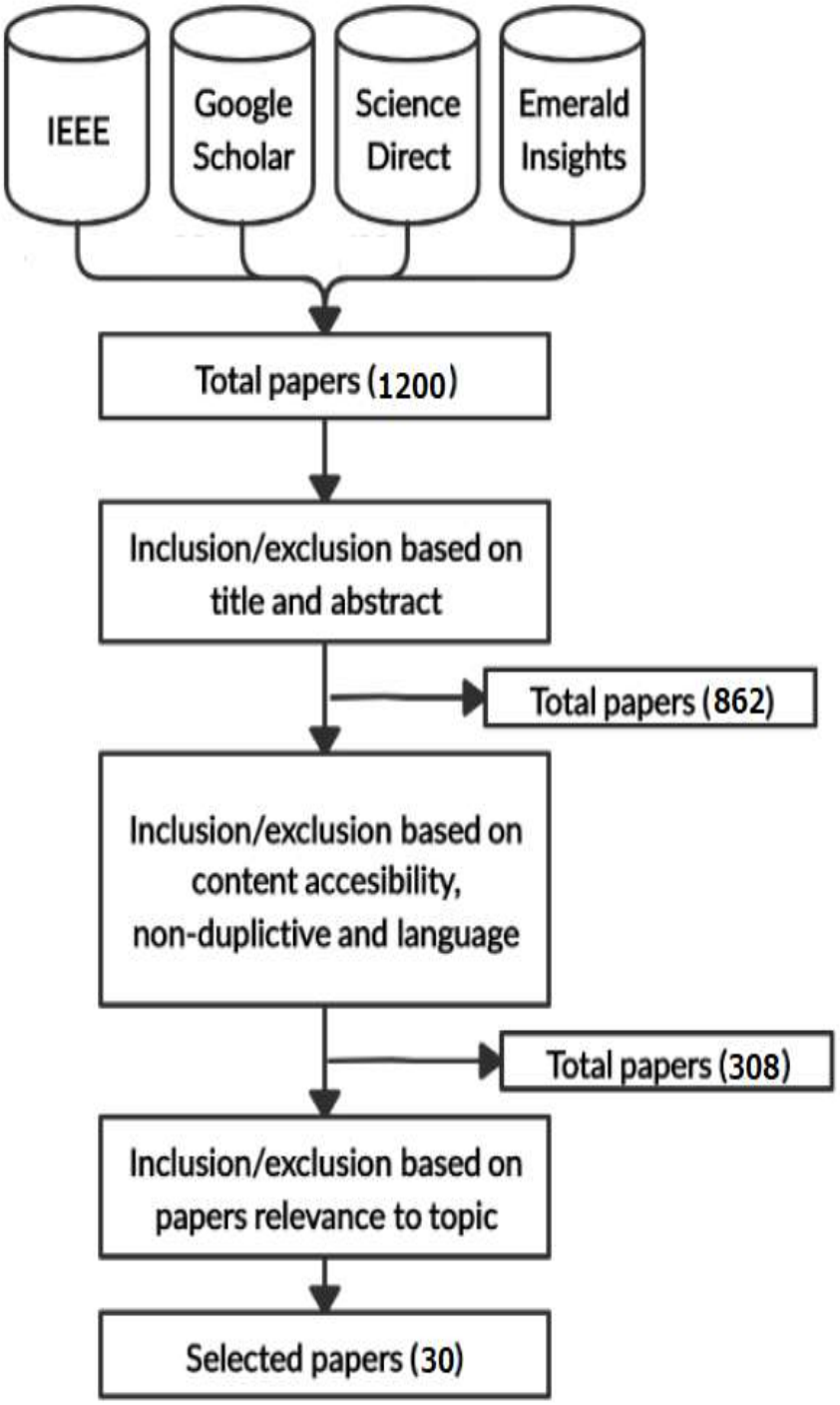


FIGURE 3.2: Systematic Literature Review for the research study

3.1 Modelling COVID-19 trends using multiple Machine Learning Techniques

Shah et al. (2020) utilized a linear regression model to predict the number of cases in Pakistan. Various parameters were chosen to assess the spread of COVID-19 across the country over a span of time. A scatter plot was used to determine the relationship between various variables to establish a correlation. A correlation coefficient was used to determine the number of future cases in Pakistan (Shah, 2020). Similarly, Daniyal et al. (2020) worked on a statistical model to predict the pandemic's progress in Pakistan and the number of future death cases. The age and gender were also collected in order to better understand the future trends. Three main regression models were used. Linear regression model, logarithmic and quadratic regression models. Also, R-square, Adjusted R-square and AIC along with BIC criteria were used to assess the validity of the models. The study concluded that the modelling showed that the cases would decrease in October 2020 (Daniyal, 2020).

Moreover, Saba et al. (2021) used time-series and machine learning models for forecasting daily confirmed COVID-19 cases in ten different countries for example India, China, Iran and Pakistan. The study utilized random forests, Support Vector Machine, Decision Trees, ARIMA and polynomial regression models for this purpose. Error based performance criterion were used to assess the accuracy of each model. The models were however, unable to produce all possible trends due to the varying nature of the data sets and the lockdown types employed (Saba, 2021).

In addition to this, [1] utilized the ARIMA (Autoregressive Integrated Moving Average) model along with multiple statistical software were utilized to predict the cumulative cases in Pakistan. The root means square and means absolute error was used to determine the accuracy of the model. This research concluded that ARIMA was able to predict precisely the cases in the next ten days.

Table 3.1: Review Matrix for Literature Review

Authors	**Proposed technique**	**Limitations**	**Evaluation**	**The dataset used**

Tanzila Saba, Ibrahim Abunadi, Mirza Naveed Shahzad and Amjad Rehman Khan	**Random forests, K-nearest neighbors, SVM, DTs (Decision Trees), polynomial regression, Holt winter, ARIMA, and SARIMA. The accuracy and effectiveness of the model were checked using errors based on different performance criteria.**	**The datasets varied depending on the size, nature, and type of lockdown so it was difficult to pinpoint the optimal model.**	**A detailed evaluation was carried out in Python using mean absolute percentage error, mean absolute error and root mean square error. The models: Holt's winter, ARIMA, and SARIMA produced optimal results. Herd lockdown policies were the best.**	**The data set was collected for nine different countries from January 2020 till September 2020.**

Muhamad Daniyal, Roseline Oluwaseun Ogundokun, Khadijah Abid, Muhammad Danyal Khan and Opeyemi Eyitayo Ogundokun	**Three different regression models are used: linear, logistic, and quadratic.**	**The models do not take into account the number of active cases and recoveries.**	**Three regression models were chosen and the quadratic modeling.**	**The dataset is obtained from the NIH (National Institute of Health), Pakistan from February 2020 to August 2020.**
Muhamad Ali, Dost Muhammad Khan, Muhammad Aamir, Umair Khalil and Zardad Khan	**The study will utilize Autoregressive Integrated Moving Average for forecasting the cumulative cases.**	**ARIMA cannot perform well if the trend in the data is not upwards and linear. Instead, Autoregressive conditional heteroscedastic is suggested.**	**RMSE and MAE were utilized for checking model accuracy. ARIMA s a better forecasting model.**	**Data is taken from the website of the Ministry of National Health Service of Pakistan from 27th February 2020 to 24th June 2020.**

Syed Tahir Ali Shah, Abeer Iftikhar, Muhammad Imran Khan, Majad Mansoor, Adeel Feroz Mirza and Muhammad Bilal	**A Linear Regression model is utilized along with the correlation coefficient.**	**The models do not take into account the number of active cases and recoveries.**	**Supervised Learning was used for this. Polynomial regression is used for analysis purposes**	**Daily WHO updates and official government sites are used for data collection from January 2020 to September 2020.**
Aman Khakharia, Vruddhi Shah, Sankalp Jain, Jash Shah, Amanshu Tiwari, Pratamesh Daphal, Mahesh Warang and Ninad Mehendale	**Use of ARMA, ARIMA, Linear Regression, Bayesian Ridge Polynomial Regressor, Support Vector Regressor, Random Forest, XGBoost,**	**The codes for the percentage errors run for a range of predicted numbers for all ML models and not for the actual predictions made by each model. No single standard model is achieved for all countries.**	**The accuracy of each model was checked.**	**Data were collected from 10 different countries including Pakistan, India, Bangladesh, etc.**

Iftikhar Ahmad and Syed Muhammad Asad	**Artificial Neural Network with rectifying linear unit-based technique.**	**More realistic input from real life can be used to produce effective controlling parameters for decision-making.**	**Error for the ANN model is checked using the mean absolute error.**	**The dataset is from 25th February 2020 to 10th July 2020. It is collected from the CSSE and ESRI along with Pakistan's official site**
M. Yousaf, Samiha Zahir, Muhammad Riaz, Sardar Muhammad Hussainb and Kamal Shah	**ARIMA is used.**	**Uncertainty of optimal time of virus disappearance**	**Akaike information criterion is used.**	**The dataset is obtained from NIH, Pakistan.**
G.D. Barmparis and G.P. Tsironis	**The Gaussian fitting hypothesis is used.**	**Each data is different from the other so difficult to observe the effectiveness of the model.**	**Mean error is used for evaluation.**	**Dataset is taken from official sites of various countries.**
Peipei Wanga, Xinqi Zhenga, Jiayang Li, Bangren Zhua	**Logistic models are used along with the Prophet model as a hybrid model.**	**The new hybrid model was not validated**	**No error observing technique was used.**	**Dataset of Brazil, Russia, India, Peru and Indonesia is used from ArcGIS platform.**

Iman Rahimi, Fang Chen, Amir H. Gandomi	**Systematic Literature Review.**	**Less number of studies are considered here.**	**Studies compared for biases.**	**Previous literature was collected using Scopus and Web of Science databases.**
Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung Won On, Waqar Aslam and Gyu Sang ChoI	**Logistic Regression, SVM, least absolute shrinkage, and exponential smoothing is used for this purpose.**	**Difficult to use SVM as difficult to create an accurate hyperplane. Need to do on an updated dataset along with real-life forecasting.**	**R-squared score, Adjusted R-square score, mean square error, and mean absolute error along with root mean square error is used.**	**GitHub and CSSE are used to collect dataset.**

Sweeti Sah, Surendiran, R, Dhanalakshmi, Sachi Nandan Mohanty, Fayadh Alenezi,and Kemal Polat	**Prophet, ARIMA, and Hybrid Stacked LSTM-GRU models are utilized**	**Each model was different, so it becomes extremely difficult to observe the effectiveness of each model**	**LSTM and GRU models outperformed compared to all different predictive models, with regard to R square and RMSE.**	**WHO, Kaggle, Johns Hopkins Hospital, GitHub, China CDC, Conference of State Bank Supervisors, DataHub, Italy Ministry of Health, COVID-19 Radiography Database, U.S. National Institutes of Health and Georgia State University's Panacea Lab are utilized to obtain dataset.**

Zohair Malki, El-Sayed Atlam, Ashraf Ewis, Guesh Dagnew, Ahmad Reda Alzighaibi, Ghada EL-marhomy, Mostafa A. El-hosseini, Aboul Ella Hassanien, and Ibrahim Gad	**SARIMA is utilized.**	**The model was not flexible, robust and resilient to manage unpredictable future events and situations.**	**The precision of the developed SARIMA model was checked using the normal distribution principle.**	**WHO and Johns Hopkins University official websites are used to acquire datasets.**
Sarbhan Singh, Bala Murali Sundram, Kamesh Rajendran, Kian Boon, Tahir Aris, Hishamshah Ibrahim, Sarat Chandra Dass, and Balvinder Singh Gil	**ARIMA is used.**	**Smoothened data and independent covariates could have been used to improve the accuracy of the ARIMA model**	**The root means square and means absolute error was utilized to decide the precision of the ARIMA model.**	**The dataset was obtained from the official MOH (Ministry of Health) of Malaysia, as well as John Hopkins University websites.**

Ambreen Chaudhry, Aamer Ikram, Muazam Abbas, Mumtaz Ali Khan, Tayyab Rathore, Moin Iqbal, Nosheen Awan, Akram Qamar, Sana Abbasi, Shafique Rehman, Tamkeen Ghafoor, Mirza Amir Baig, and Jameel Ahmed Ansari	**ARIMA is utilized.**	**The models do not take into consideration the pandemic preparedness and response and response strategies at regional and district levels**	**The predicted cases demonstrated a stationary exponential growth with a 95 percent confidence level for the next 60 days.**	**The dataset was acquired from the official website of Pakistan.**

Ram Kumar Singh, Meenu Rani, Akshaya Srikanth Bhagavathula, Ranjit Sah, Alfonso J Rodriguez-Morales, Himangshu Kalita, Chintan Nanda, Shashi Sharma, Yagya Datt Sharma, Ali A Rabaan, Jamaica Rahmani, and Pavan Kumar	**ARIMA Model is used.**	**The model did not support any volatility changes during the prediction time period.**	**Its value was modeled with 95 percent, 80 percent, and 70 percent confidence intervals, as well as the 95 percent confidence intervals were represented as the median interval between the 80 percent and 70 percent values.**	**The data was obtained from Worldometer.**

Considering the significant challenges during the Coronavirus outbreak, to mitigate the impacts of this epidemic, the improvement of compelling prediction models decidedly affected successfully accurate estimates of future trends. Additionally, these modeling techniques permit doctors and health managers to create strategic arrangements for any future variability in disease trends. Direct forecasting models and artificial intelligence (AI) approaches end up being effective techniques for estimating Coronavirus cases and different infectious illnesses [1]. The upsides of machine learning and AI for time series forecasting and modeling lie in the adaptability of learning dynamic behaviour of data, complexity, and nonlinearities, for example, as observed in epidemiological information such as those in infectious

diseases. Various regression models, along with neural networks, have been successful in predicting various diseases in patients and help provide an early response for effective healthcare [2].

3.2 Predictions and Modelling of Coronavirus cases utilizing the ARIMA Model

Different time series models are used in forecasting; a few of them found in the literature are AR, MA, ARMA, SARIMA, AFRIMA, ARIMA and non-linear models together with ARCH, GARCH etc. Nevertheless, these models are utilized in situations where more data is required, as referred to in the work of F. Khan, A. Saeed and S. Ali[12]. Additionally, the ARIMA model was predict in the work of M. Aslam [13], Mumtaz T.A Aslam. F [14], Yousaf et al. [15] to anticipate Coronavirus cases in Pakistan. Also, the ARIMA model has widely been utilized for predicting in various countries, such as by Tandon et al. [9], Perone.G [16], and D. Benvenuto et al. [17]. The Autoregressive Integrated Moving Average model is introduced for predicting affirmed cases. To formulate a Autoregressive Integrated Moving Average model, a total of 41 days of data were collected. In countries like Italy, China, South Africa, Iran, and Thailand, they used the Autoregressive Integrated Moving Average model to create two kinds of graphs: ACF (Autocorrelation Function Graph) and PACF (Partial Autocorrelation Graph). Lastly, a predictive analysis of affirmed cases was shown. During and after the lockdown was implemented, the Corona impact was considered, emphasizing on just the positive instances, as well as the Prophet model, a time-series analytic model with excellent performance in real-time data, was utilized. They found out that positive examples increase in a directed way during the lockdown time period, in comparison to, the relaxation time period, and they contended that the lockdown was legitimate with severe guidelines that might prevent the spread of Coronavirus in developing nations, such as India [10]. Coronavirus new and total deaths were projected and afterwards analyzed using various models. Time-series prediction models, like Autoregressive Integrated Moving Average, Prophet, and Seasonal Autoregressive Integrated Moving Average models were utilized to predict. The most suitable models were created utilizing the minimum values of RMSE (Root Mean Square Error), MAE (Mean Absolute Error), MSE (Mean Square Error), MAPE (Mean Absolute Percentage Error), and AIC (Akaike Information Criterion). The model's fit was assessed, and projections for the subsequent two to

three weeks were produced to be used in the future for these time-series models [11-13]. An Autoregressive Integrated Moving Average model was utilized to anticipate Coronavirus cases on the basis of Johns Hopkins epidemiology data. Between January 2020 and February 2020, the Coronavirus dataset is accessible. An A.R. (Autoregression), a M.A. (Moving Average), as well as a Seasonal Autoregressive Integrated Moving Average model make up the Autoregressive Integrated Moving Average model. Differences were favored to stabilize the time-series log conversion. The Autocorrelation Function and Partial Autocorrelation Function correlogram were utilized to test the Autoregressive Integrated Moving Average model. The Autoregressive Integrated Moving Average model performed better, compared to other models in predicting Coronavirus cases, based on the results.

3.3 Predictions and Modelling of Coronavirus cases utilizing the Logistic Growth Model

To model the positive cases, research scholars have utilized different growth models, such as Logistic regression, ARIMA, SI lR, and SEIR models, for their work. Because of the unpredictable nature of the virus, SI IR and SEIR models have been considered to be ineffective. ARIMA and logistic growth models have been deemed to be more effective by numerous researchers. Logistic growth models have been widely used to predict and model COVID-19 cases. The reason behind this is the simple principle of efficient calculations. Consequently, Wang et al. [3] have used a logistic growth model, therefore, predicted outbreaks for nations, like global, Brazil, Russia, India, Peru and Indonesia to estimate the cases in future. Meanwhile, Apiano F. Morais [4] investigated the use of logistic regression for fitting data from different countries such as China, Iran, Italy, South Korea, Spain and the United States. Furthermore, the work done for Saudi Arabia has developed ARIMA and Logistic growth models to identify the trends and provide forecasts of COVID-19 cases as investigated in the work of Tusneem Elhassan and Ameera Gaafar [5]. Additionally, F. Rojas et al. [6] in their work, calculated three models: Logistic, Gompertz, and SIR to model and forecast Coronavirus cases. In addition to this, Chen et al. ‘s work has utilized a five-parameter logistic growth model to fit the data available and make predictions for COVID-19 [7]. Christopher Y. Shen [8], in the study, has investigated the use of a logistic growth model whose parameters are calculated using the NLS (non-linear least square) technique. Furthermore, Farhan Saif [9] has used Discrete Generalized Logistic Growth to model and predict daily infections for the next few days in Pakistan.

Tatrai and Zoltan Varallyay [10], in their work, have used a logistic growth model to model predict the Coronavirus positive cases in 114 countries of the world, including Pakistan. Likewise, in work done by Rida Ahmed and Sana Ahmed [11], the use of the logistic growth model has been investigated to measure the accurate number of Coronavirus cases in Pakistan to help suggest measures for effective control.

3.4 COVID-19 situation in Pakistan

As Covid-19 cases are increasing on daily basis, different research scholars have utilized time series models to make predictions. Furthermore, governments of different areas have taken precautionary measures to stop the spread of novel virus named as Covid-19. However, some researchers have also suggested some measures which governments have adopted to break the spread. This section, therefore, covers some of the suggestion found in literature.

To deal with the current pandemic situation & major lockdown situation, J Willan et al. [33] have highlighted the issues which need attention. They highlighted the issues which could be faced by hospitals such as beds & critical care capacity, isolation of Covid-19 patients, scheduling online appointments, courier services for delivery of medication, stop or pause chemotherapy sessions. Consequently, they suggested their government to appoint good project manager to handle pandemic. Adding more for the health care professionals, Adams and Walls. [34] in their work have suggested the use of mask, googles and gloves would be advantageous for the safety of workers. The alcohol-based disinfectants should be used for cleaning the instruments such as stethoscope. More in the same study, suggestions were made for the emergency care in hospitals. The new patient who has arrived in emergency should be given mask, gloves and hand care hygiene, a suspected patient should be isolated, or 6 ft. should be maintained. Frequent sessions of information and feedbacks for the workers should also be arranged. Similarly, Xie et al. [35] advised the extra care for high risk patient. Similarly, Phua et al. [36] in their study have made recommendations for intensive care unit of hospitals and the workers. More, Qiu et al. [37] made suggestions for the people who are facing psychological problems due to the lock down situation.

However, Templeton et al. [38] in their have made recommendations on the Structural inequalities in this pandemic situation. As given in work of Atchison et al. [39] the sources of information for people are Televisions, newspaper, web pages, social media, friends & family etc. The survey reports the safety measured taken

by the people such as washing hands, avoid travelling, shopping, social events, the use of face mask and hand sanitizers etc. Furthermore, Fagherazzi et al. [40] highlighted the use of social media to spread information regarding Covid-19 and highlighted the challenges. Consequently, to manage the aged people the recommendations were made in work of Giacomo et al. [40] Furthermore, the hospitals as mentioned in work of Waris A et al [41] some hospital and isolations wards are designated for providing the desirable health care to covid-19 patients so that all other people remain uninfected. Moreover, the lockdown regions of Punjab are identified and reported in work of Saeed U et al [42] .Likewise, in the study of Saif. [43], some recommendations were made for Pakistan such as the doctors, social media, educationalist should take part in spreading the information. Further, Elsheikh A et al [44] highlighted the impact of the social distancing and precautionary measures to be taken all over the country.

Since the start of the pandemic, the primary influx of the illness has been very questionable because of the obscure idea of the infection, its instrument of infectivity, transmission, and conceivable treatment choices. Nonetheless, the fast reaction of each nation, including Pakistan, controlled the deadly spread of Coronavirus. On 26th February 2020, the very first case of Coronavirus was substantiated by the Ministry of Health, Pakistan, and afterward, a persistent spreading of this disease was seen throughout the nation. This virus originally entered Pakistan through immigrants who were returning from Iran, Saudi Arabia [31], and from Pakistanis who were trapped in different nations that were brought back to Pakistan on special flights [32]. In 2020, the nation was testing more than 5000 individuals every day. Later on, each of the provinces worked toward increasing the number of tests they conducted in a day to 30,000 by the end of June 2020. As stated by NCOC (National Command and Operation Centre), the positivity rate during the initial Coronavirus wave ranged between 18 to 23 percent. The initial COVID-19 wave reached its peak on 14th June 2020 and the number of positive cases began to decrease subsequently. By 30th August 2020, the nation was conducting just 18,015 tests to confirm 208 positive cases. Nonetheless, soon the number of positive cases began to rise again and towards the end of November 2020, the National Command and Operation Centre announced the second wave of the Coronavirus. The measures taken by the government are shown below:

Date	Activity	Region	Category	Total Cases
Feb 26	1st Two Case in Pakistan in Karachi & Islamabad	Nationwide		2
Mar 4	Closed Chaman Border with Afghanistan	Nationwide	Border	5
Mar 4	Screening at Airports: Lahore, Karachi, Islamabad & Peshawar	Nationwide	Air Travel	5
Mar 6	1st Case Recovered in Karachi	Sindh		6
Mar 7	Border reopened with Taftan, Iran	Nationwide	Border	6
Mar 10	Temporary Ban in Sindh on Marriage Halls, Lawns	Sindh		19
Mar 12	Remaining Matches of PSL without crowd in Karachi	Sindh	Cricket	21
Mar 13	Stopped all international flights, except those at Islamabad, Karachi & Lahore airports	Nationwide	Air Travel	28
Mar 13	Educational Institutes to be closed till April 5 (Fed Gov)	Nationwide	Education Institutes	28
Mar 13	Educational Institutes of Sindh **closed till March 30** (Sindh Gov) later extended	Sindh	Education Institutes	28
Mar 13	Educational Institutes of Sindh **closed till March 30** (KP Gov) later extended	KP	Education Institutes	28
Mar 13	Educational Institutes of Sindh c**losed till March 30** (GB Gov) later extended	GB	Education Institutes	28
Mar 14	Educational Institutes of Sindh **closed till April 6** (AJK Gov) later extended	AJK	Education Institutes	33
Mar 16	All Border with Iran, Afghanistan & China were closed for 2 weeks	Nationwide	Border	187
Mar 17	PSL Playoffs Postponned	Nationwide	Cricket	247
Mar 18	First 2 Deaths Reported in Pakistan	Nationwide		307
Mar 19	Ban on Public Transport in Balochistan	Balochistan		447
Mar 21	All International Flights were suspended for 2 Weeks	Nationwide	Air Travel	645
Mar 22	Ban on Intercity Transport in KP	KP	Public Transport	776
Mar 22	Ban on Intercity Transport in GB	GB	Public Transport	776

Mar 23	Lockdown imposed in Sindh (Everything Shut down) **(Mar 23 to 7 Apr)**	Sindh	Lockdown	875
Mar 23	Lockdown imposed in AJK **(Mar 23 to 7 Apr)**	AJK	Lockdown	875
Mar 24	Lockdown imposed in Punjab **(Mar 24 to 6 Apr)**	Punjab	Lockdown	990
Mar 24	Lockdown imposed in Balochistan **(Mar 24 to 6 Apr)**	Balochistan	Lockdown	990
Mar 25	Lockdown imposed in Islamabad **(Mar 24 to 6 Apr)**	Islamabad	Lockdown	1049
Mar 25	Ban on Intercity Transport in ICT	Islamabad	Public Transport	1078
Mar 27	Public Holidays in KP extended to 5th April	KP		1369
Mar 28	8 AM to 5 PM Lockdown in Sindh	Sindh	Lockdown	1500
Apr 2	Lockdown extended to April 14th (All regions)	Nationwide	Lockdown	2419
Apr 10	All Flights suspended to April 21	Nationwide	Air Travel	4781
Apr 14	Educational Institutes **closed till 31 May Nationwide (All Regions)**	Nationwide	Education Institutes	5976
Apr 15	Lockdown extended for another 2 weeks	Nationwide	Lockdown	6426
Apr 24	Lockdown extended till 9th May	Nationwide	Lockdown	11940
May 9	Lockdown was eased in Pakistan (Ramadan)	Nationwide	Lockdown	29465
May 9	Construction Industry Open, Shops in Rural Areas Open	Nationwide	Lockdown	29465
May 9	Educational Insititutes **closed till 15th July Nationwide (All Regions)**	Nationwide	Education Institutes	29465
May 22-27	Eid ul Fitr Holidays			
Jun 1	Lockdown was lightened in Pakistan (Only Sat & Sun Lockdown) - Smart Lockdown	Nationwide	Lockdown	80576
Jun 1	Train Transport & Tourism in GB & KP Allowed	Nationwide	Lockdown	80576
Jun 1	All Industries Open except few	Nationwide	Lockdown	80576
Jun 22	Border with Afghanistan opened for Trade after 3 Months	Nationwide	Border	185034

Jul 10	Educational Institutes **closed till 15th Sep Nationwide (All Regions) in Phases**	Nationwide	Education Institutes	246351
Jul 31-2 Aug	Eid ul Azha Holidays			
Sep 15	Educational Institutes **opened on Sep 23rd**	Nationwide	Education Institutes	303089
Nov 16	Compliance of Strict SOPs in commercial areas (Mircro Smart Lockdown)	Nationwide	Lockdown	361082
Nov 16	Ban on Political Gathering, Wedding Guests limited to 300 **until Jan 31**	Nationwide	Lockdown	361082
Nov 23	Virtual Classes & Vacations of All educational institutes **(26th Nov to 10th January)**	Nationwide	Education Institutes	379883
Jan 15 2021	Class 9-12 Classes opened on Jan 18 and rest on **Feb 1**	Nationwide	Education Institutes	516770
Feb 25 2021	Educational Insitutes shall be open 5 days (Mon - Fri) from **March 1**	Nationwide	Education Institutes	577482
Feb 26 2021	All restrictions lifted except for opening of educational institutes	Nationwide		578679
Feb 27 2021	Vaccine dosage administrations started privately	Nationwide		579843
March 16 2021	Sinopham, Sinovac and Astrazeneca Vaccine administered publicly	Nationwide		612307
March 28 2021	Pzifer vaccine available for dosage.	Nationwide		658993
April 11 2021	All sindh educational institutes remain open while others still closed down.	Sindh	Education Institutes	725558
April 11 2021	Islamabad, Lahore, Multan, Rawalpindi, Faisalabad, Bahawalpur, Hyderabad, Peshawar, Swat and Muzaffarabad under strict lockdown.	Nationwide		725558

May 20 2021	Outdoor dining reopens	Nationwide		895205
May 24 2021	Ban on shrines, cinemas, festivals to continue	Nationwide		906669
June 1 2021	Outdoor marriage events, elective surgeries allowed	Nationwide		925316
June 21 2021	Board exams to be held	Nationwide	Education Institutes	949963
July 1 2021	Lockdowns on Saturday and Sunday only for Islamabad	ICT		960174
July 1 2021	Lockdowns on Friday and Sunday in Karachi	Sindh		960174
July 10 2021	Schools reopen in Punjab and Islamabad after summer break	Nationwide		973938
July 21 2021 - July 26 2021	Eid ul Adha holidays with complete lockdown	Nationwide		1012426
July 27 2021	Most schools closed till 31st August 2021	Nationwide		1016498
August 26 2021	Cambridge board exams nationwide	Nationwide	Education Institutes	1145295
September 1 2021	All schools reopen	Nationwide	Education Institutes	1168705
September 4 2021	Schools closed down from 5th September till 11th September	Nationwide	Education Institutes	1176472
September 4 2021	Intercity travel banned for all provinces.	Nationwide	Public Transport	1176472
October 2021 - August 2022	All commercial and personal activities resumed	Nationwide	Public	1551614

Towards the end of March 2021, 1,024,737 tests in total were conducted with 672,851 (approximately 66 percent) confirmed Coronavirus cases and over 14,590 (roughly 2.1 percent) casualties were reported. Pakistan also experienced 605,265 patients in recovery (roughly 90 percent) and the number of patients that were going under treatment from May 2020 to 31st March 2021 (Figures 3.4 and 3.5). As per the available information, around 250 healthcare professionals in Pakistan are among the fatal victims of Coronavirus. The regional government of Sindh was the first to implement a complete lockdown, because of which the spread of this virus was decreased to some degree.

Because of its outbreak, the initial wave of the Coronavirus was extremely unpredictable because of the obscure nature of the infection, its methods of infectivity, transmission as well as conceivable treatment choices. Nonetheless, the quick response of each nation, comprising Pakistan,

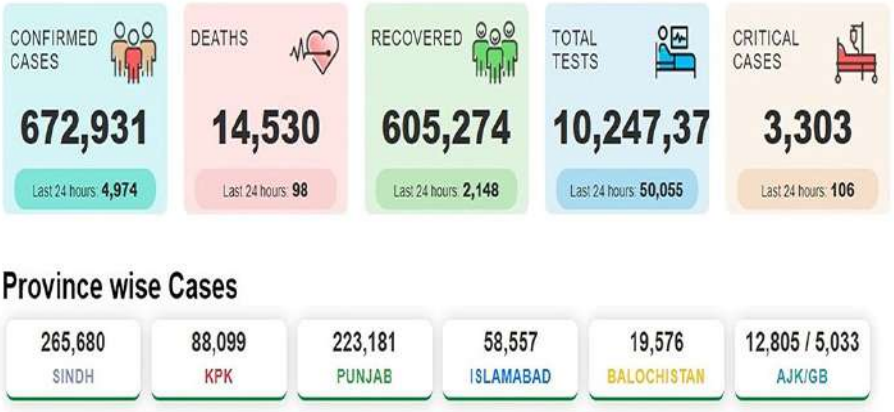


FIGURE 3.3: Accuracy graph of the proposed model

controlled the deadly occurrence of Coronavirus. During the first wave, Pakistan was exceptionally effective in dealing with the transmission of the virus. It was made conceivable because of the implementation of smart lockdown approaches. In a smart lockdown, restricted time for movement was permitted to residents after which free movement was strictly forbidden. Explicit regions inside a country were sealed, where Coronavirus positive cases were identified. Nonetheless, the second wave of the Coronavirus happened, which was middle in its pathogenicity as well as its transmission. It might have been because of the advanced development in vaccines and treatment. The World Health Organization had already alerted the government of Pakistan that the number of individuals infected with Coronavirus was predicted to surpass 0.25 million towards the end of July 2020. Nonetheless, no such anticipated infection rates were reported. Another variation of SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) arose from the UK [34] and had been detected in more than 64 nations on 27th January 2021, comprising Pakistan. This particular variation is related to an increased risk of casualty, in comparison to, different variations with average casualties of 100 patients reported daily in Pakistan. Because of the occurrence of this new variation, the 10 metropolitan areas of Pakistan, Bahawalpur, Faisalabad, Hyderabad, Islamabad, Lahore, Multan, Muzaffarabad, Peshawar, Rawalpindi, and Swat were put under strict lockdown till 11th April 2021, where the regional government was involved to observe the adoption of SOPs (Standard Operating Procedures). One thing that was common between the initial and third waves was the onset time, for example, Spring. It might result in the generation of an assumption that pollens play a substantial part in transmitting SARS-CoV-2 virus.

During the first three waves, Pakistan was exceptionally effective in making serious decisions to handle the situation created due to the transmission of the sickness. The spread was controlled using lockdown arrangements. During the lockdown, restricted time for commercial activity was permitted to residents, after which complete movement was completely restricted. In a smart or micro lockdown, explicit regions inside a city were closed off where Coronavirus-positive cases were distinguished. Pakistan has endured a total of five waves. However, the Pakistani government was still more effective compared to others in the region. Strict adherence to SOPs, along with closures of educational institutes, entertainment sites and non-essential areas, were effective in decreasing the positivity rates in many regions. Free and public vaccination campaigns also ensure that the concept of herd immunity can be taken into consideration in the large

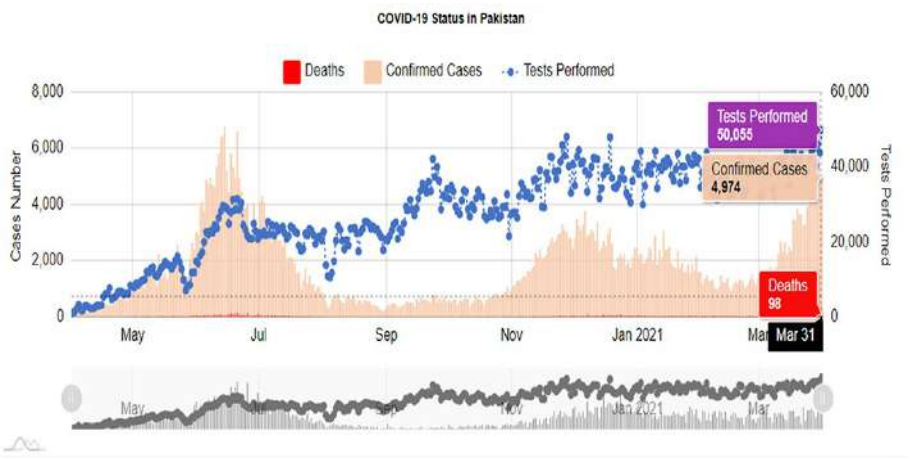


FIGURE 3.4: Possible solution approaches for predicting COVID-19 trends across various countries

Provinces	Population	Confirmed cases	Active cases	Casualties	Recoveries
Azad Jammu and Kashmir	4,038,000	1,823	587	38	1,256
Baluchistan	11,342,737	11,458	1,943	133	9,271
Gilgit Baltistan	2,130,470	1,968	345	42	1,481
Islamabad	2,000,468	13,987	2,515	159	12,019
Kyber Phakhtun Khuwa	34,518,838	33,431	5,899	1,146	24,493
Punjab	109,128,576	91,927	23,397	2,186	77,536
Sindh	58,994,355	113,215	18,877	2,152	95,398

FIGURE 3.5: Cases reported in Pakistan

population as well. In any case, multiple waves are still expected, with low to moderate symptoms after transmission. Even though the number of issues related to Coronavirus is showing an overall decrease, there are still daily infection cases throughout the region [13].

3.5 The socioeconomic effects of the COVID-19

Financial disruption related to the Coronavirus pandemic had serious effects on different financial markets. Significant events comprised the Russia-Saudi Arabia oil cost war that lead to the

collapse of oil prices as well as the stock market towards the end of March 2020. The UNDP (United Nations Development Program) anticipates a 218 billion dollars reduction in revenues in emerging nations predicted that the Coronavirus' financial effect to last for quite a long time or even years. A few financial professionals also anticipated a decrease in global Gross Domestic Product and population growth rate. As specified by the World Bank, the Coronavirus has spread with disturbing or alarming speed, infecting millions of individuals and carrying financial activity to a nearby stop as countries imposed strict restrictions on development to stop the virus from spreading. As the human cost grows, the monetary loss is currently and addresses the greatest financial shock the world has faced in ages. The June 2020 Worldwide Financial Possibilities depicts both, the short and long-term viewpoints on the effect of Coronavirus and the long-lasting destruction it has on the chances for advancement. The benchmark measure presumed a 4.9 percent withdrawal in worldwide Gross Domestic Product in 2020, using market transition standard loads, the most extensive global recession in several years, despite the remarkable efforts of governments to encounter the downturn with economic and money-related arrangement aid. Over the past few years, the significant downturns initiated by the epidemic are depended upon to leave scars that might last for a decade through lower investments, deterioration of human resources due to loss of work and education, and fragmentation of global supply and trade relations. This crisis features the prerequisite for a crucial action to pad the epidemic's prosperity and financial results, secure weak populations and set up for lasting recovery. For emerging business industries as well as developing countries, a substantial number of individuals who faced overwhelming shortcomings because of the spread of the Coronavirus, it is fundamental to support public well-being structures, resolve the problems introduced by casualness, and comprehensive changes that would reinforce strong and manageable developments after the well-being crises start to decline.

3.6 Outcomes of Literature Review

The research study examines the use of two growth models, the Logistic Growth Model and ARIMA, for the modeling of positive Coronavirus cases. The literature review carried out in this section shows that both models are popular when modeling predictive trends of daily cases of infectious diseases. The performance of these models is used to map out a timeline for the daily infection caseload for multiple regions and to determine which model gave the best results for the case of Pakistan. The performance of both (the Logistic Growth Model ARIMA) models is evaluated using Root Mean Squared Error.

3.7 Chapter Summary

This chapter presented past studies related to using ARIMA and the Logistic Growth model to predict the number of positive Coronavirus cases. We further highlighted the Coronavirus situation in Pakistan, and the socioeconomic impacts of COVID-19. The next chapter reviews the methodology used for this study and the data sources involved.

Chapter 4

Research Methodology

The study examines the use of two growth models, the Logistic Growth Model and ARIMA, for the modeling of positive Coronavirus cases. The literature review carried out in the previous section shows that both models are popular when modeling predictive trends of daily cases of infectious diseases. The performance of these models is used to map out a timeline for the daily infection caseload for multiple regions and to determine which model gave the best results for the case of Pakistan. The performance of both (the Logistic Growth Model and ARIMA) models is evaluated using Root Mean Squared Error. The ARIMA model is the most famous predicting technique, which allows for the identification, estimation, and diagnostics of the data as well. Similarly, the Logistic growth model is another popular technique which can determine the growth rate patterns for infectious diseases like COVID-19.

4.1 Research Setting

The research has been conducted in Islamabad, Pakistan. This experimentation is performed on Del Intel® Core™ i5-5200U CPU @ 2.20GHz machine with 8 GB random access memory, 64-bit operating system, and x64-based processor. Further, we performed an analysis on Python by installing the required packages and related libraries.

4.2 Research Duration

The study was conducted from February 2020 till August 2022. We empirically tested and developed the growth models by fitting the above functions to available data from all over Pakistan during the five waves of infections from 2020 to 2022. The regions include ICT (Islamabad Capital Territory), Sindh, Punjab, Baluchistan, KPK (Kyber Phakhtun Khuwa), AJK (Azad Jammu and Kashmir), and GB (Gilgit Baltistan). The study is longitudinal type where the data collected over the time period stated is used to understand and infer caseload trends.

4.3 Methods Used

The ARIMA model is the most famous predicting technique, which allows for the identification, estimation, and diagnostics of the data as well. Similarly, the Logistic growth model is another popular technique which can determine the growth rate patterns for infectious diseases like COVID-19. Figure 4.2 represents the workflow utilized for the analysis of Coronavirus data using the Logistic Growth Model and ARIMA, respectively.

4.4 Sample Size and Technique

The size of sample used for this study include the total population of Pakistan. The daily positive COVID-19 caseload is taken into consideration. A flow of the sampling strategy is given in Figure 4.2. Systematic Sampling is used to zero down on the desired population the research questions will be addressing. It is a probability sampling method where the study has chosen the elements required from the target population i.e the daily number of cases for COVID-19 from

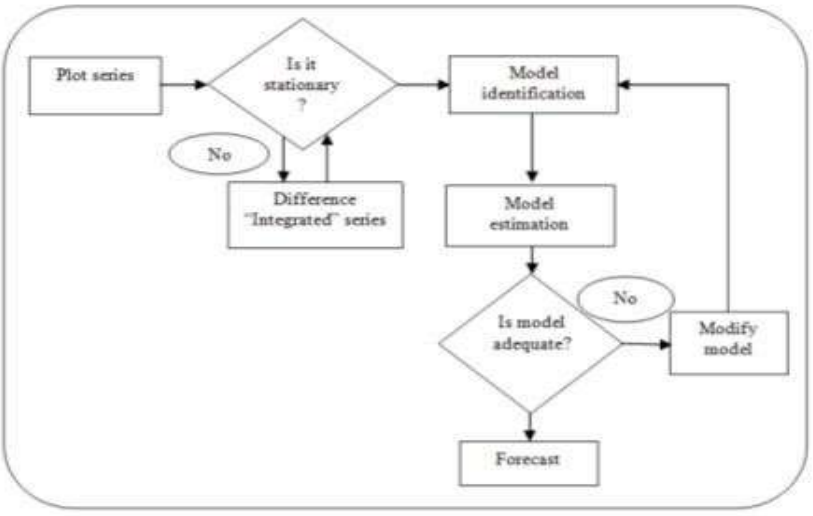


FIGURE 4.1: Workflow for methods utilized in study

different regions of Pakistan. The regions include ICT (Islamabad Capital Territory), Sindh, Punjab, Baluchistan, KPK (Kyber Phakhtun Khuwa), AJK (Azad Jammu and Kashmir), and GB (Gilgit Baltistan).

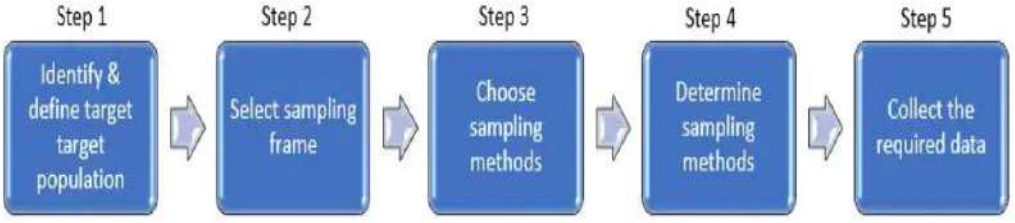


FIGURE 4.2: Sampling Strategy

4.5 Data Analysis

4.5.1 Logistic Growth Model

The Logistic growth model was first used by LotKa [18] and later by [24] in 1998 for population dynamics. However, Logistic growth models can be utilized in the situation where the increasing growth rate is seen at the start; however, later the growth rate starts to decrease, as mentioned in the work of [5]. Logistic growth models have been used extensively to model trends for COVID-19. Moreover, it has been successful in modeling disease trends for other infectious diseases, such as Ebola and Dengue successfully. We used the following equation to map out the COVID-19 cases onto the logistic growth model:

$$\frac{1}{C}\frac{dC}{dt} = r(1\frac{C}{K}) \tag{4.1}$$

```
def my_logistic(t, a, b, c):
  return c /(1 + a* np.exp(-b*t) )
p0 = np.random.exponential(size=3)
p0
bounds = (0, [10000., 3., 220000])
import scipy.optimize as optim
X = np.array(df_first_wave_Pakistan['Day'] +1)
y= np.array(df_first_wave_Pakistan[area])

(a, b, c), cov = optim.curve_fit(my_logistic, X, y, bounds=bounds, p0=p0)
a , b, c
def logistic(t):
  return c /(1 + a* np.exp(-b*t) )
plt.figure(figsize=(6,4))
plt.scatter(X,y, marker='.', color='red')
plt.plot(X, logistic(X), linewidth=2)
plt.title('Logistic Model FIT on ' +area + ' Data of Covid-19')
plt.legend(['Logistic Model', 'Real Data'])
plt.xlabel('Time')
plt.ylabel('Infections')
plt.xticks(range(1,190,30),["15-Mar","15-Apr","15-May","15-Jun"], rotation=20);
z = np.array(range(215))
plt.figure(figsize=(6,4))
plt.scatter(X,y, marker='.', color='red')
plt.plot(z, logistic(z))
plt.title('Logistic Model FIT and Prediction for ' +area+ ' over Next 30 Days')
plt.legend(['Logistic Model', 'Real Data'])
plt.xlabel('Time')
plt.ylabel('Infections')

plt.xticks(range(5,250,30),["15-Mar","15-Apr","15-May","15-Jun"], rotation=20);
preds = logistic(X)
actual = df_first_wave[area]
```

FIGURE 4.3: LGM Algorithm

Where C represents the number of cases. $\frac{dC}{dt}$ shows the growth rate, while r and K are the constants and Q is used to plot the S-shaped curve of the Logistic equation. The S-shaped curve of the Logistic equation for the individual waves can be used. The Logistic Growth Model Algorithm is shown in Figure 4.3

4.5.2 ARIMA Model

ARIMA models were introduced by Box and Jenkins in 1970, as stated in the work of [21] and is widely used for making short-term forecasting. Consequently, the aim of using a time series model such as ARIMA is to predict the number of Coronavirus-positive cases. However, after some interval, the time series become stationary then the ARMA model is referred to as Autoregressive Integrated Moving Average (ARIMA) model, as mentioned in the work of [17]. Further, the ARIMA model applies to seasonal and non-seasonal forecasting. To test if the data is stationary or not, the Augmented Dickey-Fuller (ADF) unit root test will be utilized:

$$\Delta Z_t = \alpha_o + \gamma Z_{t-1} + \alpha_2 t + \sum_{i=1}^{k} \beta_i \Delta Z_{t-i} + \varepsilon_t \tag{4.2}$$

The equation of regression above is utilized to test if the time series is stationary or not. Null hypothesis H_o: Z_t is nonstationary & alternative hypothesis H_a: Z_t is stationarity tested by the following equation of regression. Refrain from disapproving H_o if the p-value is more compared to the standard level of significance (which is usually 0.05) which concludes that the given time series is nonstationary; otherwise, do not reject H_1 [14].

To assure that the time series remains stationary, logarithmic transformations along with differences will be considered. The autocorrelation function (ACF) and partial autocorrelation function (PACF) are used to estimate the parameters (p, d, q) of the ARIMA model.

As given in the work of Mumtaz T.A Aslam. F [14], the ARIMA model is definite as three order parameters p, d, and q, where p corresponds to the periods taken for the autoregressive model, d is the order of integration, and q denotes the periods taken in the moving average.

When we use AR, a linear model, to model yt, we made subsequent calculations:

$$y_t = \ \delta + \ \phi_1 y_{t-1} + \ \phi_2 y_{t-2} + .. \ + \ \phi_p y_{t-p} + \ \epsilon_t \tag{4.3}$$

where is the intercept term, *yt − i are regressors, ϕt-i is and ϵ signifies the error term (ϵϵ).* MA is an alternative class of linear model. In MA, the target variable is modelled via its own incorrectly predicted values of present and preceding times. It can be written as follows in terms of error terms:

$$y_t = \ \mu + \ \theta_1 \epsilon_{t-1} + \ \theta_2 \epsilon_{t-2} + .. \ + \ \theta_q \epsilon_{t-q} + \ \epsilon_t \tag{4.4}$$

The mathematical form of ARMA (p, q) is as follows:

$$y_t = \{\ \delta + \ \phi_1 y_{t-1} + \ \phi_2 y_{t-2} + .. \ + \ \phi_p y_{t-p} + \ \epsilon_t\} + \{\mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + .. + \theta_q \epsilon_{t-q} + \epsilon_t\} \quad (4.5)$$

Precisely, we can amend the above equation as follows:

$$y_t = \delta + \sum_{i=1}^{p} \phi_i y_i + \sum_{j-1}^{q} \theta_j \epsilon_{t-j} \ + \ \epsilon_t \quad (4.6)$$

The algorithms below were used in the Jupyter notebook in Figure 4.4 and 4.5:

In the next step, we empirically tested and developed the above models by fitting the above functions to available data from all over Pakistan during the five waves of infections from 2020 to 2022. The regions include: Islamabad Capital Territory (ICT), Sindh, Punjab, Baluchistan, Kyber Phakhtun Khuwa (KPK), Azad Jammu and Kashmir (AJK) and Gilgit Baltistan (GB).

4.6 Dataset(s)

The data for this study is collected from the daily reported cases from the official website (https://covid.gov.pk/stats/pakistan) of the government of Pakistan. The website is updated daily regarding information on COVID-19 cases. The data is collected from 03/10/2020 to 01/08/2022. We collected data for 874 days. We have performed the analysis using the confirmed number of cases reported every day for Pakistan, the federal capital and the provinces. The regions include: Islamabad (ICT), Azad Jammu and Kashmir (AJK), Baluchistan, Gilgit Baltistan (GB), Khyber Pakhtunkhwa (KPK), Punjab and Sindh. The data prepossessing algorithm is given below in Figure 4.6:

4.7 Evaluation method(s) and criteria

There are four primary targets of time series analysis.

4.7.1 Model Evaluation

The Root Mean squared error is used to evaluate the performance of the Logistic Growth Model and ARIMA model for the Prediction of COVID-19 cases. The data set is composed of accumulated cases from March 10th 2020 to August 01st 2022. Equation 6 shows the mathematical calculations for Root Mean Squared Error.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} p_t^2} \quad (4.7)$$

Where n is considered as a time point sequence, p_t is the difference between actual and predicted values at a given time t. The lower values for RMSE would indicate the better performance of the model.

```
import numpy as np                                          Importing required libraries
import matplotlib.pyplot as plt
import pandas as pd

df=pd.read_excel('Location of dataset')                     Loading required dataset

df.head()                                                   Observing first 5 rows of
df.dropna(inplace=True)                                     dataset

                                                            Removing all rows with null
df.columns = ['day','patients']                             values

df.shape                                                    Creating two columns

                                                            Finding number of rows and
y= df.set_index('day')                                      columns of dataset

                                                            Setting index value to day
y.plot(figsize=(15, 6))                                     column

                                                            Plotting the given dataset
plt.show()

from statsmodels.tsa.arima_model import                     Importing library for ARIMA
ARIMA

from statsmodels.graphics.tsaplots import                   Importing library for
plot_acf,plot_pacf                                          acf(Autocorrelation
                                                            Function) and p
```

```
from statsmodels.tsa.arima_model import                     Importing library for ARIMA
ARIMA

from statsmodels.graphics.tsaplots import                   Importing library for
plot_acf,plot_pacf                                          acf(Autocorrelation
                                                            Function) and p
                                                            acf(Partial Autocorrelation
                                                            Function) plots

plot_acf(y)                                                 Plotting acf which helps to
```

FIGURE 4.4: ARIMA Algorithm

4.7.2 Description

Data is described using compact statistics and graphical process. A time series graph in which data is plotted over a time period is explicitly more useful.

	decide p parameter
plot_pacf(y)	Plotting pacf which helps to decide q parameter
y_train=y[:(len(y)-5)] y_test=y[(len(y)-5):] y_test	Making train and test datasets
patient_model=ARIMA(y_train,order=(p,d,q)) patient_model_fit=patient_model.fit() patients_forecast=patient_model_fit.forecast(steps=5)[0] patients_forecast	Fitting ARIMA model on suitable values of p,q and d and forecasting next five values
from statsmodels.tsa.holtwinters import ExponentialSmoothing train=y[:(len(y)-5)]	Importing Exponential Smoothing and creating another train dataset

FIGURE 4.5: ARIMA Algorithm (Continued)

Code	Comment
print(data.shape)	Printing the shape of dataset
print(data.dtypes)	Printing the data-types of all columns of dataset
print(data.info())	Information of dataset like column names,non-null count,dtypes of all columns
print(data.describe())	Description of dataset like mean,max,etc
print(data.isna().sum())	Checking for missing values in the dataset(null values)

FIGURE 4.6: Data Pre-Processing

4.7.3 Modeling

One of the main purpose of the time series analysis is to design a suitable analytical model which could be used to explain the data. It is substantial to review the univariate model of a variable depended just on its previous values, whereas a multivariate model just not depends on the past values of the understudy factor. However, also relies on the present and past values of different

regressors. In the last scenario, the change in one time series could help to describe or portray the fluctuations in different time series.

4.7.4 Predicting

When an individual observes a time series, it might require estimating or computing the next values of that particular time series. Predicting is a substantial application of time series and might help in eradicating issues in different circumstances.

4.7.5 Control

The primary objective of conducting a time series analysis is to control the quality of a provided procedure, this procedure might belong to any arena. The most important feature of a time series analysis is to fit a statistical model to the provided series for which upcoming values are to be estimated.

4.7.6 Chapter Summary

In this chapter, we have discussed the setting of the research, the duration of the research, the Machine Learning methods that are used in the research, and the different evaluation methods and criteria that would be used to evaluate the accuracy of the Machine Learning methods. The next chapter reviews the proposed framework that was used in the research.

Chapter 5

Proposed Solution for modeling COVID-19 trends in Pakistan

A time series analysis is observed after dividing the data into the following movements: secular trend, seasonal variations, cyclical movement and unpredictable movement. The paper examines the use of two growth models, the Logistic Growth Model and ARIMA, for the modeling of positive Coronavirus cases. The literature review carried out in the previous section shows that both models are popular when modeling predictive trends of daily cases of infectious diseases. The performance of these models is used to map out a timeline for the daily infection caseload for multiple regions and to determine which model gave the best results for the case of Pakistan. The performance of both (the Logistic Growth Model and ARIMA) models is evaluated using Root Mean Squared Error.

5.1 Proposed Solution

This study examines the use of two growth models Logistic Growth Model and ARIMA for prediction of positive cases of COVID-19. The performance of both (Logistic Growth Model and ARIMA) models is evaluated using Root Mean Squared Error as shown in Figure 5.1.

5.1.1 Secular trend movement

The direction of variations within a time series might be either, in a lower or higher direction for a specific time frame. These fluctuations are primarily deemed as long-term fluctuations. These fluctuations might be because of various factors, like the information about the capital, a growth in population, an enormous scale shift or change in customer demands, and so on. For example, an increase in price and population over several years is seen. These are the two examples of upward movement. In case, new items are launched into a new market, their prices are extremely high and the sales of the items might be decreased. This is an example of downward movement. Declining and upward movement fluctuations of a time series are referred to as "Secular movements".

5.1.2 Seasonal movement

Typically, seasonal changes do not last for a long period of time and occur occasionally, once or twice a year. These changes proceed to happen every year. Significant factors because of which seasonal changes occur are atmospheric conditions and public requirements. It is generally observed seasonal fluctuation, such as clothing items that are made up of woollen are sold more in the winter season, while, on the contrary, the sales of frozen yoghurt are high in the summer and much lower in the winter.

5.1.3 Cyclic movement

Cyclical variations are the long-term fluctuations which show up in lower or higher directions in a time series. However, the cycle's duration is over one year. These fluctuations are not considered as steady or constant. One example of a cyclic fluctuation is the ups and downs in business operations. A business cycle represents these movements in four stages: time of recession, time of success, time of gloom and period of recovery. A business cycle is completed

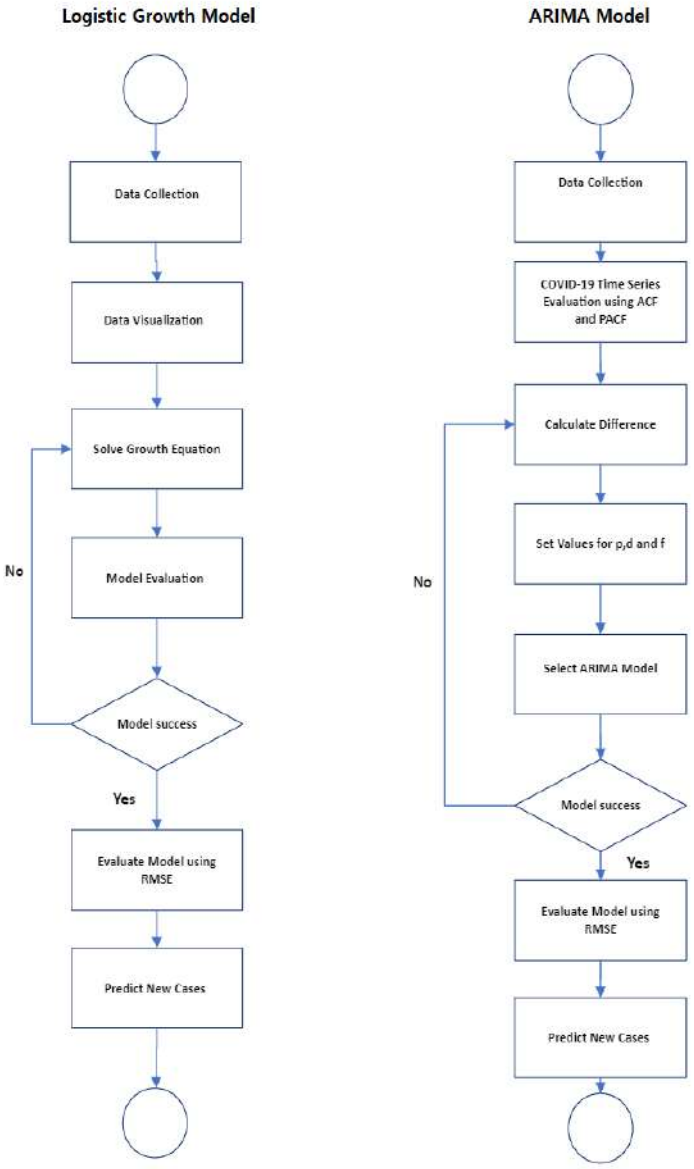


FIGURE 5.1: Proposed Solution

when these stages pass from passes starting with one and then onto the next in this particular order.

5.1.4 Unpredictable movement

These variations are short-term fluctuations and are undependable. These fluctuations are also referred to as coincidental variations as they address that aspect of fluctuations, which is not deemed inside the time series, resulting in secular patterns, cyclical fluctuations and seasonal fluctuations. Irregular fluctuations allude to events that cannot be predicted easily, like wars, volcanic eruptions, earthquakes, and so on.

5.1.5 Stationary or fixed time series

One of the major concepts of time series is Stationarity. In case, data of time series variables around a constant mean and fluctuation after a while, then the data is referred to as stationary or fixed data. A time series is called strong stationary in the event that the joint distribution of Zt1, Zt2, - - -, Ztn is identical to the joint distribution of Zt1+T, Zt2+T,- - -, Ztn+T $\forall$ t1+T, t2+T, - - -,tn+T. A time series plot is typically utilized to determine the Stationarity of the data.

5.1.6 Achieving Stationarity

5.1.6.0.1 Regular Differencing It is crucial to remove non Stationarity from a time series prior to analyzing it for building a model. In case there is a pattern in the mean, then only it requires the understudy data by performing regular or standard differencing. For any data that is non-seasonal, standard differencing of order is adequate to accomplish Stationarity. The first difference is calculated using the following equation:

$$\Delta Z_t = Z_t - Z_{t-1} \tag{5.1}$$

and the second difference is computed using the following formulae:

$$\Delta\Delta Z_t = \Delta^2 Z_t = \Delta Z_t - \Delta Z_{t-1} \tag{5.2}$$

From the equations that are mentioned above, it may be noticed that the first regular difference is utilized to eliminate linear patterns, while a second regular difference is used to eliminate parabolic patterns. Overall, time series could be differenced 'd' times to achieve Stationarity.

5.1.6.0.2 Seasonal Differencing The model identification differentiates any irregularity present in the data. Moreover, it also attempts to find the order of seasonal AR as well as seasonal MA terms. A single seasonal differencing is adequate for data in which irregularity or seasonality happens. For data that is obtained monthly, seasonal AR and seasonal MA terns might be included. For the box-Jenkins method, one cannot remove seasonality before fitting the model. In the SARIMA structure, seasonal differencing is utilized on the time series prior to

producing autocorrelation and partial autocorrelation graphs. It helps when the seasonal part of a model needs to be recognized. During certain cases, seasonal differencing might eliminate a few or complete seasonality impacts. A seasonal difference operator of period "S" could be described using the following equation: A converted time series could be obtained as the figure "t" time period as well as the figure of the series "t-s" time period, while the seasonal operator is applied to the data.

$$\nabla_{\mathrm{s}} = 1 - \mathrm{B}^{\mathrm{S}} \tag{5.3}$$

Note that

$$\nabla_{\mathrm{s}} \nabla^{\mathrm{S}} = (1 - \mathrm{B})^{\mathrm{s}} \tag{5.4}$$

A converted time series could be obtained as the figure "t" time period as well as the figure of the series "t-s" time period, while the seasonal operator is applied to the data.

$$\nabla_{\mathrm{s}} \mathrm{Z_t} = (1 - \mathrm{B}^{\mathrm{s}})\, \mathrm{Z_t} = \mathrm{Z_t} - \mathrm{Z_{t-s}} \tag{5.5}$$

5.1.7 Autocorrelation and partial autocorrelation function

5.1.7.0.1 ACF (Autocorrelation Function) One fundamental approach to testing Stationarity is based on ACF (Autocorrelation Function). Fundamentally autocorrelation function computes the level of correlation between the observations of the time series. Autocorrelation is represented by utilizing the following equation:

$$\rho_{t,\ s} = Corr(Z_t, Z_h) = \frac{\mathrm{cov(Zt, Zh)}}{\sqrt{\mathrm{var\ \ (Zt)} * \mathrm{var}(Z_{\mathrm{h}})}} \qquad \forall\ t, h\ \epsilon\{0, \pm 1, \pm 2, - - -\} \tag{5.6}$$

The coefficient of the autocorrelation is calculated based on the data sample by utilizing the following equation:

$$r_k = \frac{\sum_{\mathrm{t=k+1}}^{\mathrm{n}} \left(\mathrm{Zt} - \overline{\mathrm{Z}}\right) \left(\mathrm{Z_{t-k}} - \overline{\mathrm{Z}}\right)}{\sum_{\mathrm{t=1}}^{\mathrm{n}} \left(\mathrm{Zt} - \overline{\mathrm{Z}}\ \right)^2} \tag{5.7}$$

5.1.7.0.2 The sample data distribution of coefficients of the Autocorrelation Function The autocorrelation coefficient of an unexpected time series are roughly distributed with mean zero and standard deviations $\frac{1}{sqrtn}$

5.1.7.0.3 PACF (Partial Autocorrelation Function) The correlation between Zt-k and Zt is calculated after the effect of intermediate factors has been eliminated.

5.1.7.0.4 SPACF (Sample Partial Autocorrelation Function) The SPACF for a time series is characterized using the formulae:

$$\mathrm{r}_{kk}, fork = 0, 1, 2, 3 - - \tag{5.8}$$

$$\mathrm{r00} = 1 \tag{5.9}$$

$$\mathrm{r11} = \ \mathrm{r1} \tag{5.10}$$

$$r_{kk} = \frac{r_k - \sum_{j=1}^{k=1} r_{k-1\ ,\ jr_{k-j}}}{1 - \sum_{j=1}^{k-1} r_{k-1,\ jr_j}}, k = 2, 3, 4, - \tag{5.11}$$

where $r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-1}$, j=1,2,—-,k-1, r_{kk} is an estimation of $\widehat{\emptyset}_{kk}$

5.1.8 Models of time series

5.1.8.0.1 The AR(p) (Autoregressive model) ARIMA models were introduced by Box and Jenkins in 1970, as stated in the work of [21] and is widely used for making short-term forecasting. Consequently, the aim of using a time series model such as ARIMA is to predict the number of Coronavirus-positive cases. However, after some interval, the time series become stationary then the ARMA model is referred to as Autoregressive Integrated Moving Average (ARIMA) model, as mentioned in the work of [17]. Further, the ARIMA model applies to seasonal and non-seasonal forecasting.

The equation of regression above is utilized to test if the time series is stationary or not. Null hypothesis H_o: Z_t is nonstationary & alternative hypothesis H_a: Z_t is stationarity tested by the following equation of regression. Refrain from disapproving H_o if the p-value is more compared to the standard level of significance (which is usually 0.05) which concludes that the given time series is nonstationary; otherwise, do not reject H_1 [14].

To assure that the time series remains stationary, logarithmic transformations along with differences will be considered. The autocorrelation function (ACF) and partial autocorrelation function (PACF) are used to estimate the parameters (p, d, q) of the ARIMA model.

As given in the work of Mumtaz T.A Aslam. F [14], the ARIMA model is definite as three order parameters p, d, and q, where p corresponds to the periods taken for the autoregressive model, d is the order of integration, and q denotes the periods taken in the moving average.

When we use AR, a linear model, to model yt, we made subsequent calculations:

$$y_t = \ \delta + \ \phi_1 y_{t-1} + \ \phi_2 y_{t-2} + .. \ + \ \phi_p y_{t-p} + \ \epsilon_t \tag{5.12}$$

where is the intercept term, *yt − i are regressors, ϕt-i is and ϵ signifies the error term (ϵϵ).*

5.1.8.0.2 The MA(q) (Moving Average model) MA is an alternative class of linear model. In MA, the target variable is modelled via its own incorrectly predicted values of present and preceding times. It can be written as follows in terms of error terms:

$$y_t = \ \mu + \ \theta_1 \epsilon_{t-1} + \ \theta_2 \epsilon_{t-2} + .. \ + \ \theta_q \epsilon_{t-q} + \ \epsilon_t \tag{5.13}$$

5.1.8.0.3 ARMA (Autoregressive Moving Average) model The mathematical form of ARMA (p, q) is as follows:

$$y_t = \{ \ \delta + \ \phi_1 y_{t-1} + \ \phi_2 y_{t-2} + .. \ + \ \phi_p y_{t-p} + \ \epsilon_t \} + \{ \mu + \ \theta_1 \epsilon_{t-1} + \ \theta_2 \epsilon_{t-2} + .. \ + \ \theta_q \epsilon_{t-q} + \ \epsilon_t \} \tag{5.14}$$

Precisely, we can amend the above equation as follows:

$$y_t = \delta + \sum_{i=1}^{p} \phi_i y_i + \sum_{j-1}^{q} \theta_j \epsilon_{t-j} \ + \ \epsilon_t \tag{5.15}$$

In the next step, we empirically tested and developed the above models by fitting the above functions to available data from all over Pakistan during the five waves of infections from 2020 to 2022. The regions include: Islamabad Capital Territory (ICT), Sindh, Punjab, Baluchistan, Kyber Phakhtun Khuwa (KPK), Azad Jammu and Kashmir (AJK) and Gilgit Baltistan (GB).

5.2 ARIMA (Autoregressive Integrated Moving Average) model

Autoregressive Moving Average model has a general application which is referred to as a time-series Autoregressive Integrated Average model. This particular model is utilized when the time-series data is stationary in nature. Most of the time, time-series understudy is non-stationary. In these type of situations, the series is needed to be included in order to eliminate non-stationary nature of the data, and afterward apply the Autoregressive Integrated Average model.

Autoregressive Integrated Average model in back shift notation can typically be written as:

$$(1 - \alpha_1 \mathrm{B}^1 - - - - \alpha_\mathrm{p} \mathrm{B}^\mathrm{p})(1 - B)^d Z_t^{'} = (1 + \Theta_1 \mathrm{B}^1 - - - - \Theta_\mathrm{q} \mathrm{B}^\mathrm{q})\epsilon_\mathrm{t} \tag{5.16}$$

5.2.0.0.1 Seasonal Autoregressive Integrated Moving Average model

To deal with the seasonal changes of a time-series, an Autoregressive Integrated Moving Average model can be improved. This improved form of model is known as Seasonal Autoregressive Integrated Moving Average model (p,d,q) $(\mathrm{P,D,Q})_s$.

where p, d & q are the order of non-seasonal Autoregressive, regular differencing and non-seasonal Moving Average and P, D & Q are the order of seasonal Autoregressive, seasonal differencing and the seasonal Moving Average accordingly. S is seasonality period (S=4 in quarterly data, S=12 in a year data. Sinmilar is shown in Figure 5.2.).

5.3 The Box-Jenkins Method

In econometrics, Autoregressive Integrated Moving Average is an extended form of Autoregressive Moving Average models. Box and Jenkins method typically is applied when one needs to make predictions by finding out the most appropriate fit of a time-series discovered on past values of that particular series. It is appropriate to use in situations, where data indicates non-stationary pattern and differencing is utilized again to achieve Stationarity. Employing such methods create an Autoregressive Integrated Moving Average model.

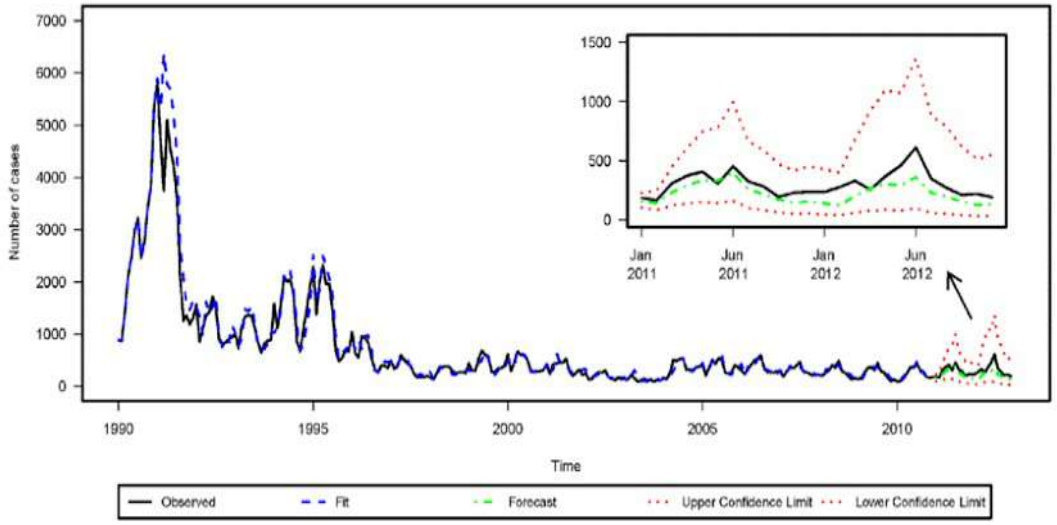


FIGURE 5.2: Seasonal Autoregressive Integrated Moving Average model

Box-Jenkins method is a robust technique for handling or managing difficult problems associated with time-series. It typically consists of four iterative process as shown in Figure 5.3:

1. Identification of the tentative
2. Estimation
3. Diagnostic testing
4. Predicting

5.3.1 Identification of the tentative

The main stage in facilitating Box Jenkins approach is the identification of the tentative. In order to use the Box and Jenkins methodology, the understudy time-series needs to stationary or fixed. In the event that the time-series is non-stationary, one needs to convert it into stationary by utilizing different technique. During this particular stage, the primary criteria utilized to determine the order of the model are Autocorrelation Function and Partial Autocorrelation Function. The Autocorrelation Function plot is usually applied to determine the order of the Moving Average model, while the Partial Autocorrelation Function is typically applied to determine the order of the Autoregressive model.

Hypothetical scheme of Autocorrelation Function and Partial Autocorrelation Function for Autoregressive Moving Average model that is for non-seasonal and Autocorrelation Function and Partial Autocorrelation Function for Autoregressive Moving Average for pure-seasonal are presented in the tables below:

5.3.2 Detection of Stationarity

A sequence plot is recognized to be utilized to check the data's Stationarity. It needs to indicate persistent location as well as scale for the stationary time-series. Particularly non-Stationarity

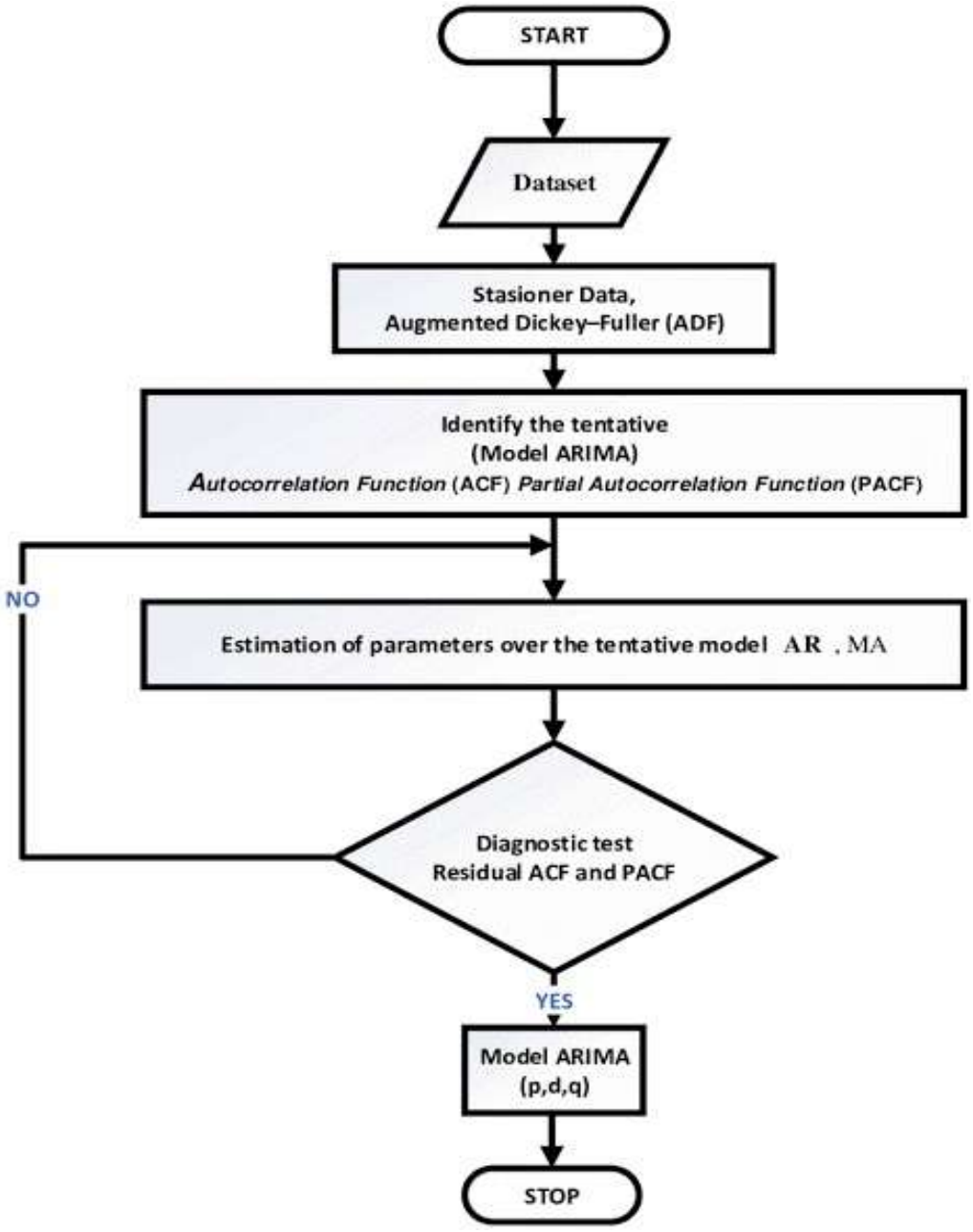


FIGURE 5.3: The Box-Jenkins Methodology

of the data is mainly indicated with an extremely slow decline or decrease in the autocorrelation graph.

5.3.3 Evaluating Stationarity of the time-series

5.3.3.0.1 ADF (Augmented Dickey Fuller) In order to check the Stationarity of a time-series, unit root tests are conducted. For this reason only, Augmented Dickey Fuller test is utilized. It is an improved form of the Dickey Fuller test. In a Dickey Fuller test, error terms are presumed that they are not correlated. In case, the errors terms are correlated and for the issue of serial correlation, Dicky Fuller introduced one more test, referred to as Augmented

Table 5.1: Hypothetical Scheme for ACF and PACF for non-seasonal time-series data

Model	Autocorrelation Function	Partial Autocorrelation Function
Autoregressive (p)	It declines in an exponential or sinusoidal manner	It discontinuous after lag p
Moving Average (q)	It discontinues after lag q	It declines in an exponential or sinusoidal manner
Non-seasonal Autoregressive Moving Average (p,q)	It declines in an exponential or sinusoidal manner	It declines in an exponential or sinusoidal manner

Table 5.2: Hypothetical Scheme for ACF and PACF for seasonal time-series data

Model	**Autocorrelation Function**	**Partial Autocorrelation Function**
Autoregressive (p)	It tails-off at lags k_s K=1, 2, 3,...	It cuts-off after lag p_s
Moving Average (q)	It cuts-off after lag Q_s	It tails-off at lags k_s K=1, 2, 3,...
Pure-seasonal Autoregressive Moving Average (p,q)	It tails-off at lags k_s K=1, 2, 3,...	It tails-off at lags k_s K=1, 2, 3,...

Dickey Fuller unit root test as shown in Figure 5.3. This specific test is utilized to manage more complicated Machine Learning models.

5.3.3.0.2 Differencing to achieve Stationarity

Box and Jenkins presented regular and seasonal differencing methods to achieve Stationarity. In case, the seasonality is strong, one needs to take seasonal differencing first. In certain situations, single differencing (either regular or seasonal) is sufficient, however, on the contrary, both differencing methods could be used as per the requirements of the data.

5.3.4 The selection criteria for choosing a model

Several criteria are accessible which can be utilized to decide a suitable Machine Learning model. In order to test, if a Machine Learning model is appropriate to use or not, two penalty functions, like AIC (Akaike Information Criteria) and BIC (Schwarz Bayesian Information Criteria) could be utilized to select the model. These functions are typically utilized to verify the adequacy of the chosen Machine Learning models. During the comparison phase, the model which has least Akaike Information Criteria and Schwarz Bayesian Information Criteria statistics are chosen as it has residuals that appear like a white noise procedure.

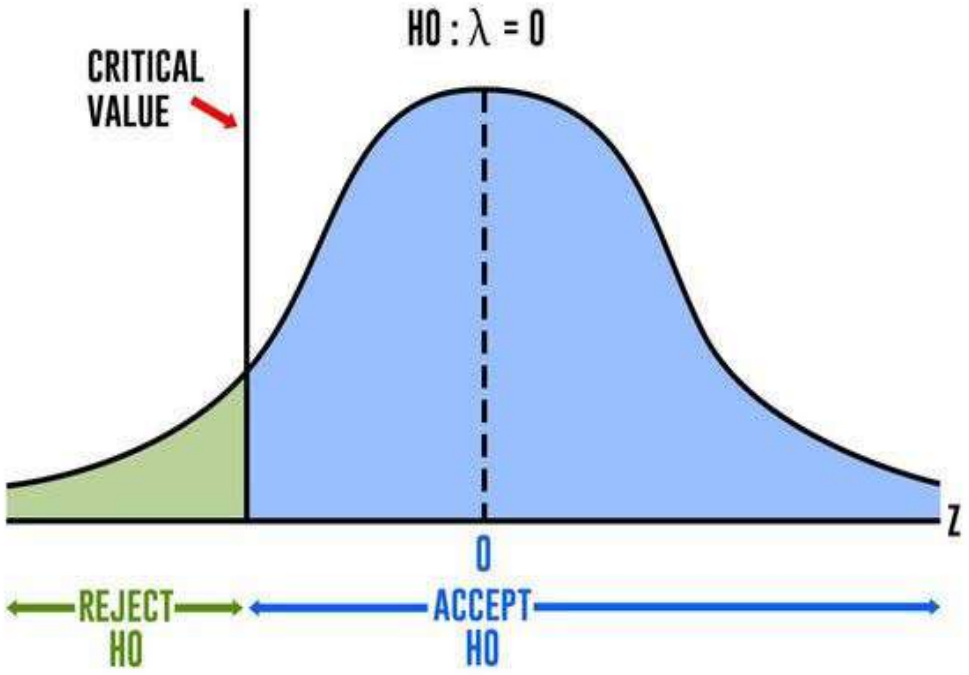


FIGURE 5.4: The Augmented Dickey Fuller test

5.3.4.0.1 AIC (Akaike Information Criteria) Akaike Information Criteria can be mathematically described using the following equation:

$$AIC = 2k - 2log(L) \tag{5.17}$$

Akaike Information Criteria is equivalent to twice the number of parameters of the statistical model substrate twice the log possibility function.

As Akaike Information Criteria is utilized as a dependable criteria for selecting a model.

5.3.4.0.2 BIC (Schwarz Bayesian Information Criteria) One more criteria or standard presented by Schwarz (1978) is Schwarz Bayesian Information Criteria, which is typically utilized in the final choice of the model and it is mathematically described using the following formula:

$$BIC = -2log(likelihoodfunction) + klog(n) \tag{5.18}$$

Schwarz Bayesian Information Criteria is equivalent to k times logarithm of no. of observations minus substrate the value of the maximized log possibility function.

Schwarz Bayesian Information Criteria is also deemed to be dependable standard in final model choice.

5.3.4.0.3 R-Square Criteria The R-square measure gives the percentage of the entire variability in the time-series that is expounded by the fitted model. A model is asserted as the best fit model when it has a maximum R-square value (maximum value is up to 1).

R^2 criterion is characterized using the following equation:

$$R_s^2 = 1 - \frac{\sum_t \left(Z_t - \widehat{Z}_t\right)^2}{\sum_t \left(Z_t - \overline{z}\right)^2}$$

Z_t is the observed time-series, $\widehat{Z}_t$ is the time-series that have fitted values and $\overline{\Delta z}$ is the simple mean that has transformed differenced series (Harvey, 1989).

5.4 Estimating the parameters of the tentative model

In box-Jenkin methods, the evaluation stage is considered the initial phase. There are several techniques, like method of moment, MLE and least square which can be utilized when one needs to estimate the parameters of a tentative model.

Typically ARIMA models are such type of nonlinear that need to use a nonlinear model method and it is frequently automatically done by a software package, like SPSS. In some software packages, the researchers have the option of the estimation process and according to that they could choice an appropriate techniques as per the stated issue.

5.5 Model diagnostics

The third stage includes building the model, that is, to diagnostically examine the model which needs to be used to make estimations. Model presumptions diagnostic checks could be analyzed statistically. The most significant and a fundamental assumption in ARIMA modeling demands that the residuals should be normally and independently distributed. Such a time-series that is serially independent is deemed as a white noise time-series. In case, diagnostic tests do not prove that the assumption is true, then the model is not suitable to use. This indicates that a few modifications is required in the model. In the event that the diagnostic tests prove the supposition to be true, then errors of the model need to be serially independent and should follow Gaussian distribution with mean zero and consistent variance. In practice, in ARIMA modeling, various models possibly fit the data, and the parameters of each model are calculated and afterward, diagnostic tests are conducted to evaluate the validity of the models. The model which fulfils the standards of various diagnostic checks is afterward utilized for making predictions. Particularly, various diagnostic checks, like residual plots of the Autocorrelation Function and Partial Autocorrelation Function, Ljung box chi-square test, and normality test are utilized for assessing the residuals arbitrariness'.

5.5.1 Autocorrelation Function and Partial Autocorrelation Function plots of residuals

Using the appropriate model to the data understudy, the Autocorrelation Function plot of residual is developed and it should have insubstantial autocorrelation at any lag. Similarly, residual Partial Autocorrelation Function plot must show clear insignificant spikes at all lags.

When no significant spikes in Autocorrelation Function and Partial Autocorrelation Function plots of the residual are encountered, then it implies that suitable fitting is in practice. Nonetheless, some spikes which are close to a significant level might happen. One could expect 1 statistically significant lag out of 20 lags by likelihood in a 95 percent confidence interval. Such spikes in residual plots might not be complicated as the position of that lag is of the problem and also determining their significance and proper judgment needs to be utilized.

5.5.2 Normality Test

In case model residuals fulfill the standards to normality presumption, the model is considered the most appropriate model to use. A histogram of model residuals grants the indication of normality. A normal probability plot is also utilized for assessing the data that is nearly normally distributed. In normal probability, graph data is utilized in such a way that it is plotted against the theoretic normal distribution that indicates a straight line.

In case the points are not close to a straight line then it declares that there is no normality.

5.5.3 Ljung Box Chi-Square Test

Ljung box promotes an advanced version of portmanteau statistics and is utilized to test the arbitrariness of the residuals. The null hypothesis is a set of autocorrelations of residuals which is referred to as a white noise during this evaluation. This test statistic is utilized to compute the significance of residual autocorrelations as the full set. In a similar way, it indicates if they are mutually significant as shown in Figure 5.5.

In this test, the null hypothesis is that the set of residual autocorrelation follows white noise and alternative hypothesis is set of residual autocorrelation is far from white noise. Test statistic χ_h^2 is compared with tabulated chi-square written as $\chi^2_{\alpha,(h-m)}$,where the significance level is equal to 0.05, h is the highest lag taken and m is the number of parameters to be estimated. The crucial region in this test is: χ^2_{h} $\chi^2_{\alpha,(\mathrm{h-m})}$, then do not reject the alternative hypothesis.

5.5.4 Jarque-Bera Test

The presumption of normality could be checked by conducting a Jarque-Bera test. This test, first of all, calculates the skewness and kurtosis. The formula that can be used to calculate Jarque-Bera is provided below:

$$\mathrm{JB} = n[\frac{S^2}{6} + \frac{(\mathrm{K}-3)^2}{24}] \sim \chi^2 \tag{5.19}$$

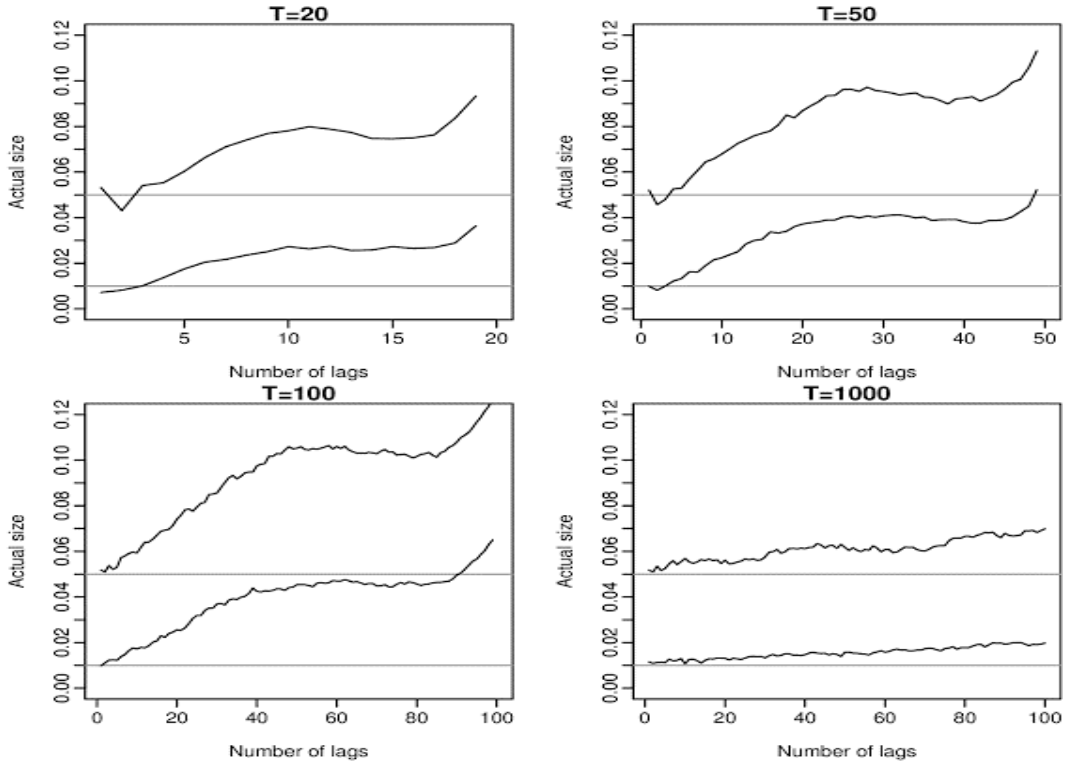


FIGURE 5.5: The Ljung box test

where n= number of observations, S= coefficient of skewness, K=coefficient of kurtosis.

A variable which is normally distributed, S=0 and K=3. Jarque-Bera evaluation of normality is typically to check the joint hypothesis that S is 0 and k is 3. In such a. Case, it is anticipated that the value of Jarque-Bera test statistic is equal to zero.

The null and alternative hypothesis in the JB test is stated as:

$H_{0:}$ Residuals are normally distributed.

H_1: Residuals are not normally distributed.

In case, the computed p-value of the Jarque-Bera test statistic is significantly low, typically, in this situation, the value of the test statistic is not equivalent to zero. Hence, it is concluded to reject H_0 otherwise do not reject H_0.

More tests other than these tests are required to test the appropriateness of the model and discover its best fit.

Two standards Akaike Information Criteria and Schwarz Bayesian Information Criteria are utilized for selecting the most appropriate fitted model. The model which has the least value of Akaike Information Criteria and Schwarz Bayesian Information Criteria is referred to as the statistically most suitable fit.

5.6 Model Forecast

The primary objective of creating the model for time-series data is to forecast the future values for understudy data. It also plays an important role in achieving prediction precision.

To acquire a minimal error forecast seven features of an ideal ARIMA model are taken into account. First of all, a model is parsimonious. Second of all, an Autoregression model need to be stationary of fixed. Third of all, a MA model should be not veritable. Fourthly, a model ought to have top-notch standard estimated coefficients. Fifth, the residuals of a model needs be uncorrelated. Sixth, a model should fit the understudy data well to satisfy the researcher. Lastly, a good fitted model needs to predict future values well means it give acceptable prediction results.

Several beneficial techniques of predicting are accessible. However, just Box-Jenkins technique is utilized. For better understanding forecasting for generalize form of ARMA(p,q) model is demonstrated.

Prior to finalizing the results, it is crucial that some prediction precision that needs to be analyzed. The most substantial forecasting precision measurements are as follows:

5.7 Measuring Forecasting Accuracy

The most commonly utilized techniques for this purpose are the MAPE (Mean Absolute Percentage Error), RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and MSE (Mean Square Error).

5.7.1 MAPE (Mean Absolute Percentage Error)

Mean Absolute Percentage Error is characterized as the mean of absolute values of the percent residuals. Mathematical equation to calculate the Mean Absolute Percentage Error is provided below:

$$MAPE = \frac{100}{\mathrm{n}} \sum |\frac{\widehat{\mathrm{Z}}_{\mathrm{t}}}{\mathrm{Z}_{\mathrm{t}}}| \tag{5.20}$$

Where sum is apply on all n absolute percent residuals and percent residuals is defines as $\frac{\widehat{\mathrm{Z}}_{\mathrm{t}}}{\mathrm{Z}_{\mathrm{t}}}$

therefore $\widehat{\mathrm{Z}}_{\mathrm{t}}$=$\mathrm{Z}_{\mathrm{t}}$-$\mathrm{F}_{\mathrm{t}}$ and Z_{t}= observed values of the time series and F_t = forecasted value. Figure 5.6 shows the MAPE equation utilization for a time-series data.

5.7.2 RMSE (Root Mean Square Error)

The equation used to calculate the root mean square error is given below:

$$RMSE = \sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left(\mathrm{Z}_{\mathrm{i}} - \widehat{\mathrm{Z}}_{\mathrm{i}}\right)^2} \tag{5.21}$$

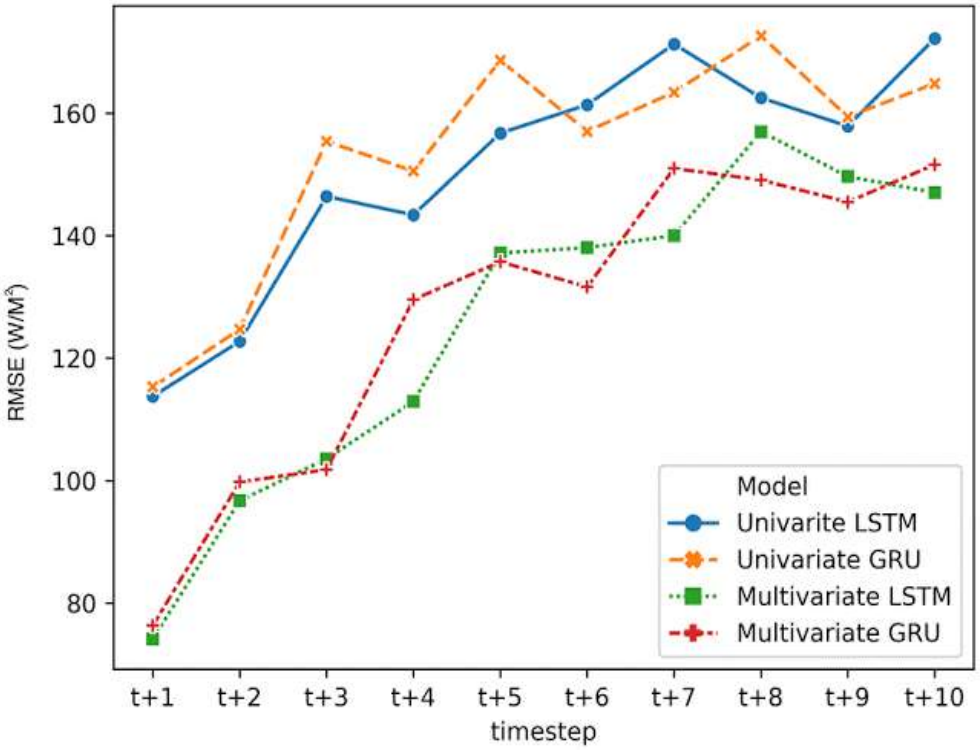


FIGURE 5.6: The Mean Absolute Percentage Error

where Z_i= actual data, $\widehat{Z}_i$= forecasted values, N = number of time period to be predicted. Figure 5.7 shows the RMSE equation utilization for a time-series data.

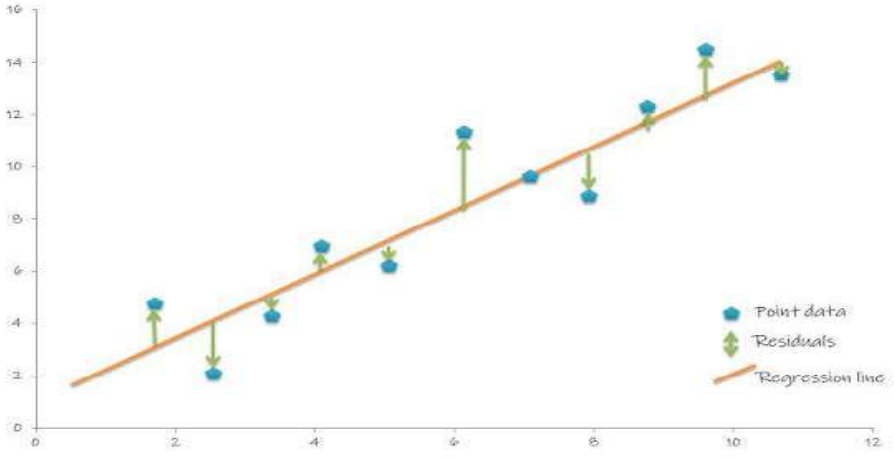


FIGURE 5.7: The Root Mean Square Error

5.7.3 MAE (Mean Absolute Error)

The formula the can be utilized to calculate MAE is given below:

$$MAE = \frac{1}{\mathrm{m}} \sum_{\mathrm{t=N-m+1}}^{\mathrm{N}} |_{\mathrm{t}} j \tag{5.22}$$

Error $(\mathrm{e_t}) is characterized as : {}_{\mathrm{t}} = \mathrm{Z_t} - \widehat{\mathrm{Z}}_{\mathrm{t}} (5.23)$ Where $\mathrm{Z_t}$ = observed value at time t, $\widehat{\mathrm{Z}}_{\mathrm{t}}$ = predicted values at time t, m is the no. of predicted values. Figure 5.8 shows the MAE equation utilization for a time-series data

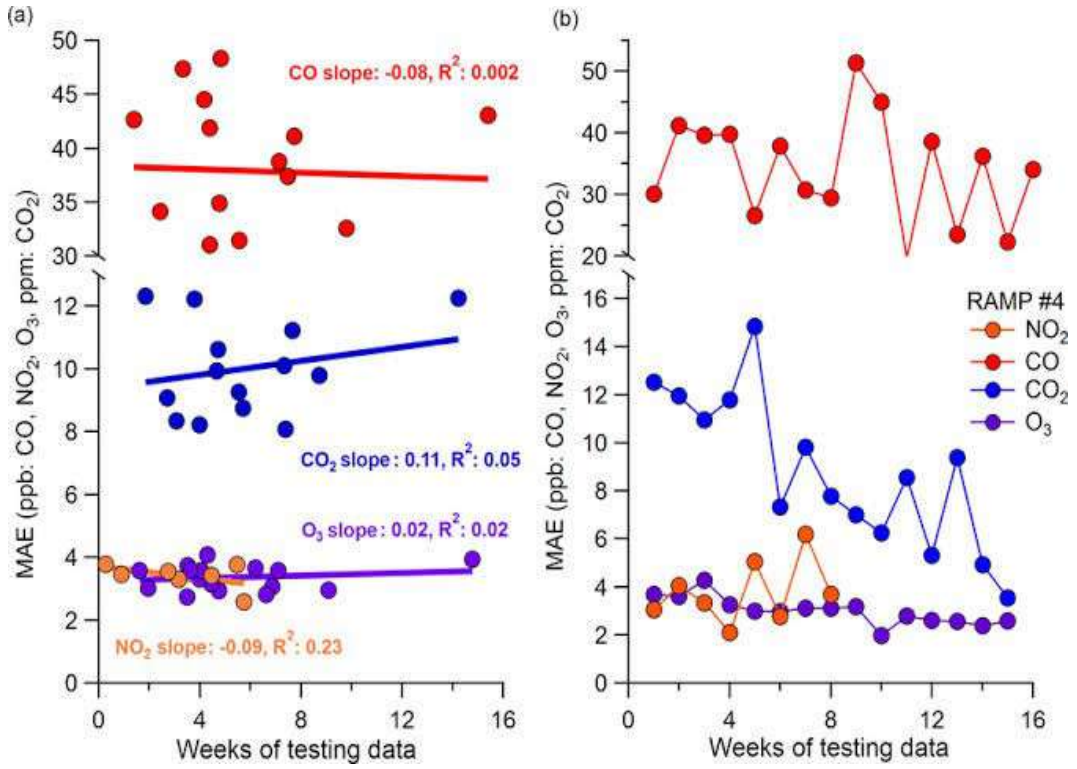


FIGURE 5.8: The Mean Absolute Error

5.8 Logistic Growth Model

The Logistic growth model was first used by LotKa [18] and later by Haberman [19] in 1998 for population dynamics. However, Logistic growth models can be utilized in the situation where the increasing growth rate is seen at the start; however, later the growth rate starts to decrease, as mentioned in the work of Wang et al [20]. Logistic growth models have been used extensively to model trends for COVID-19. Moreover, it has been successful in modeling disease trends for other infectious diseases, such as Ebola and Dengue successfully. We used the following

equation to map out the COVID-19 cases onto the logistic growth model:

$$\frac{1}{C}\frac{dC}{dt} = r\left(1 - \frac{C}{K}\right) \tag{5.24}$$

blackWhere C represents the number of cases. dC/dt shows the growth rate, while r and K are the constants. And Q is used to plot the S-shaped curve of the Logistic equation.

5.9 Chapter Summary

This chapter has introduced a framework that can be utilized to predict the number of positive Coronavirus cases accurately. It has provided detailed information about the ARIMA model that is based upon ML (Machine Learning). In the next chapter, we have explained the acquired results.

Chapter 6

Experimental Results and Evaluation

The results of COVID-19 disease modeling using the Logistic Growth Model and ARIMA model are divided into two phases. The first phase depicts the results of five Coronavirus waves. In the second phase of the results, the average caseload is studied to compare the error analysis. For daily model cases for Islamabad, Azad Jammu and Kashmir, Baluchistan, Gilgit Baltistan, Khyber Pakhtunkhwa, Punjab, and Sindh, the ARIMA model was utilized. The Logistic Growth model produced incorrect results since the model tried bringing the error to zero, so it was decided to use the model only for Pakistan's daily infection cases and for each wave separately. The rest of the session would discuss the ARIMA and Logistic Growth Model results in detail.

6.1 Dataset

The data for this study is collected from the daily reported cases from the official website (https://covid.gov.pk/stats/pakistan) of the government of Pakistan. The website is updated daily regarding information on COVID-19 cases. The data is collected from 03/10/2020 to 01/08/2022. We collected data for 874 days. We have performed the analysis using the confirmed number of cases reported every day for Pakistan, the federal capital and the provinces. The regions include: Islamabad (ICT), Azad Jammu and Kashmir (AJK), Baluchistan, Gilgit Baltistan (GB), Khyber Pakhtunkhwa (KPK), Punjab and Sindh. The waves were mapped out, as demonstrated in Table 6.1. The data was also plotted to visualize each wave and separate them out for analysis as shown below in Figures 6.1-6.8.

Table 6.1: COVID-19 Waves mapped out on the dates that occurred.

Waves	**Dates**
1st wave	**10th March 2020 to 15th September 2020**
2nd Wave	**16th September 2020 to 15th February 2021**
3rd Wave	**15th February 2021 to 25th June 2021**
4th wave	**26th June 2021 to 9th December 2021**
5th Wave	**10th December 2021 to 1st August 2022**

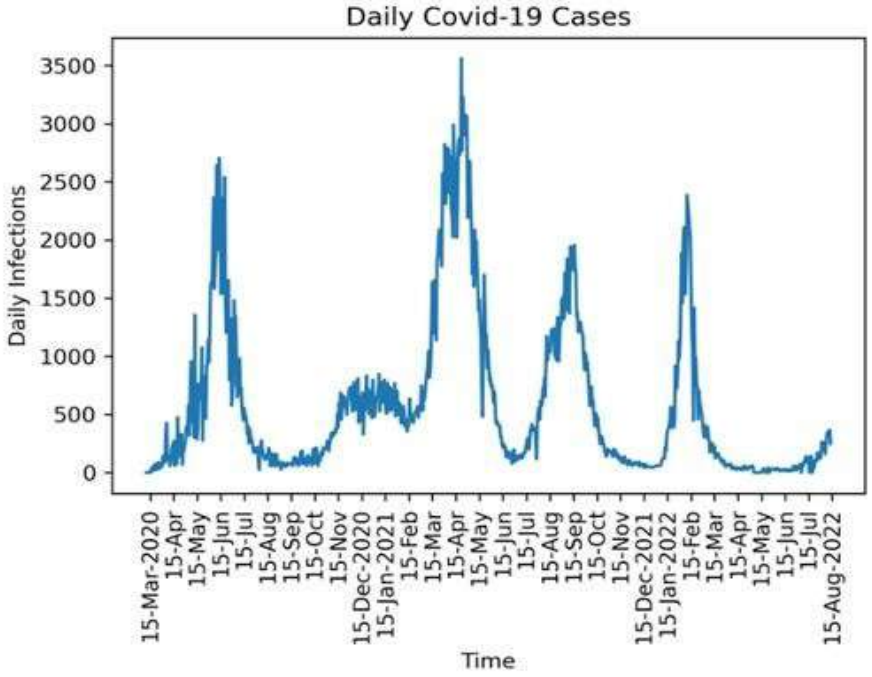


FIGURE 6.1: Visualization of daily COVID-19 cases in Pakistan

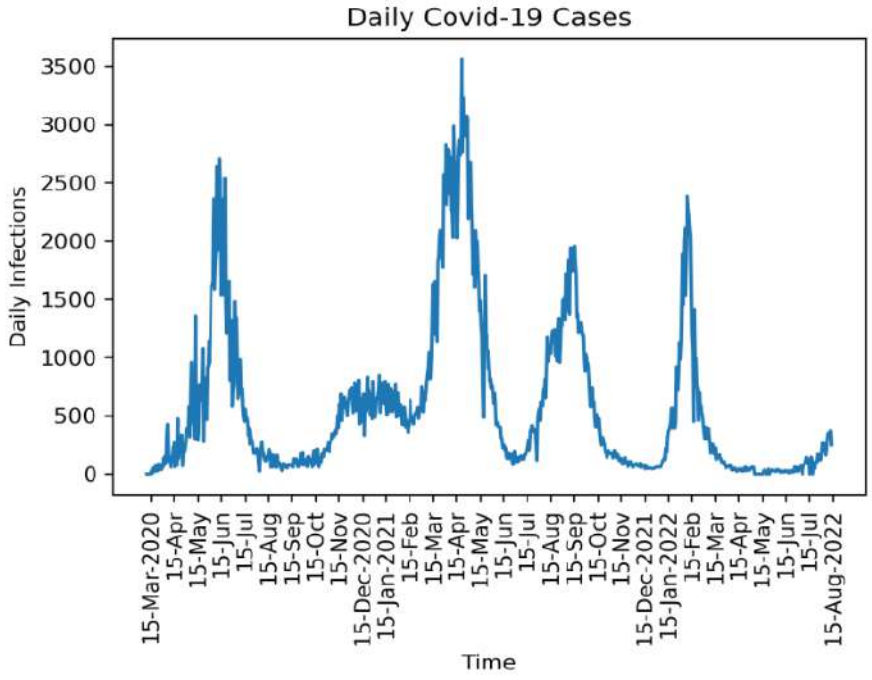


FIGURE 6.2: Visualization of daily COVID-19 cases in Punjab

6.2 Experimental Settings

This experimentation is performed on Del Intel® Core™ i5-5200U CPU @ 2.20GHz machine with 8 GB random access memory, 64-bit operating system, and x64-based processor. Further, we performed an data mining, visualization,and analysis

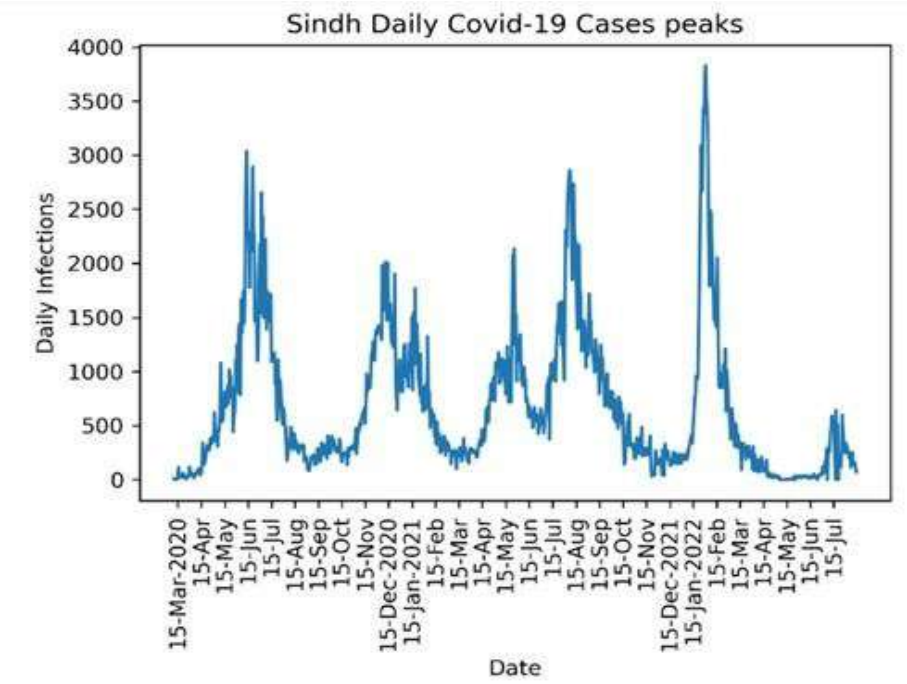


FIGURE 6.3: Visualization of daily COVID-19 cases in Sindh

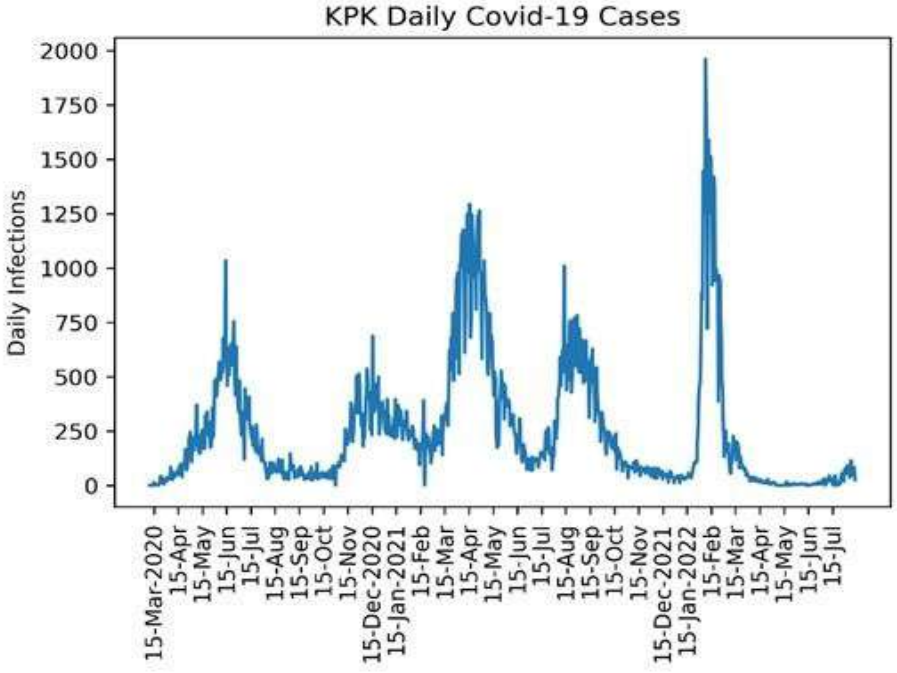


FIGURE 6.4: Visualization of daily COVID-19 cases in Khyber Pakhtunkhuwa

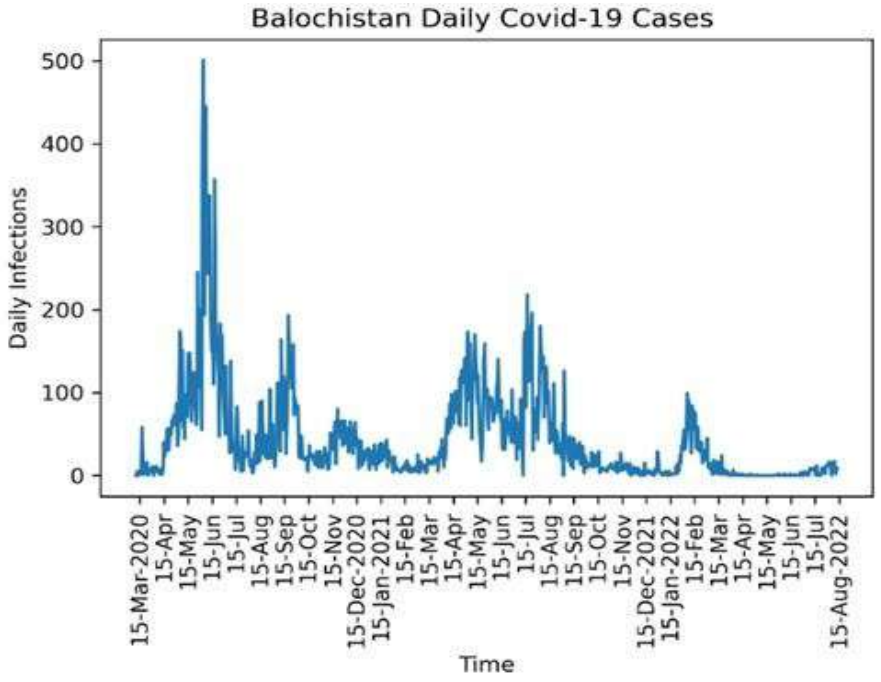


FIGURE 6.5: Visualization of daily COVID-19 cases in Baluchistan

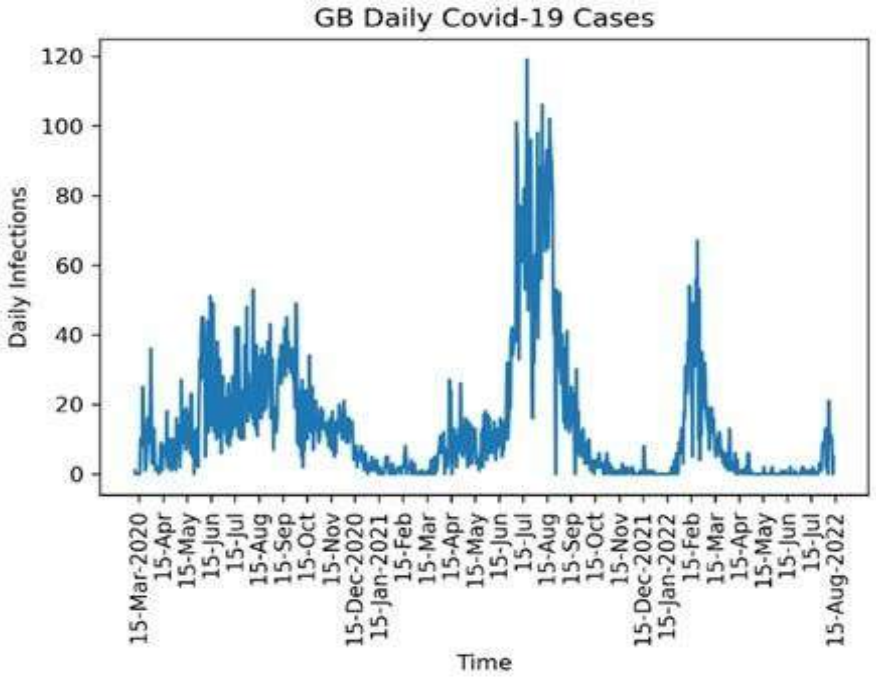


FIGURE 6.6: Visualization of daily COVID-19 cases in Gilgit Baltistan.

on Python by installing the required packages and related libraries.

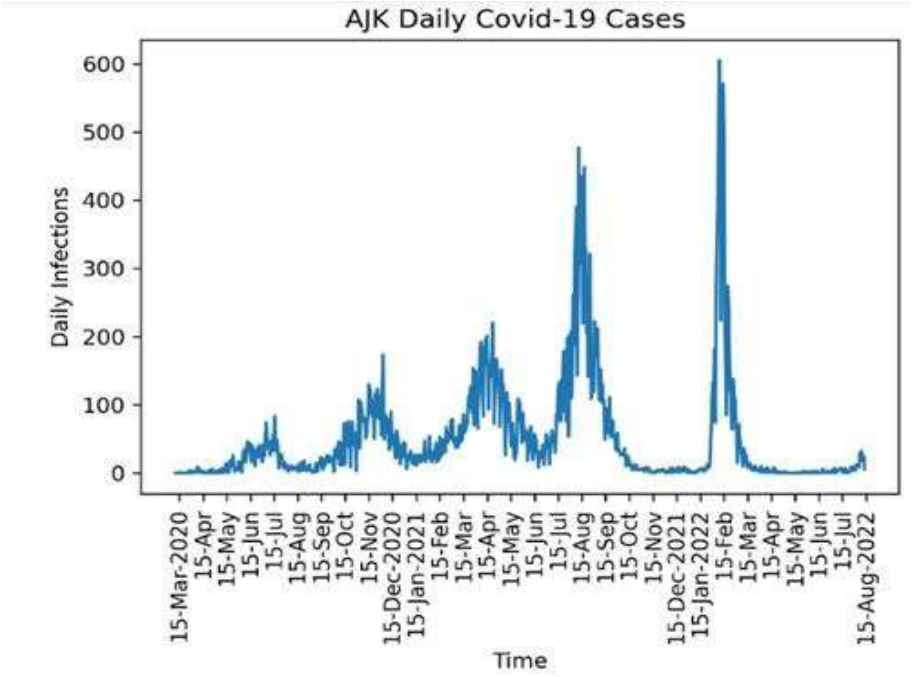


FIGURE 6.7: Visualization of daily COVID-19 cases in Azad Jammu and Kashmir
[!ht]

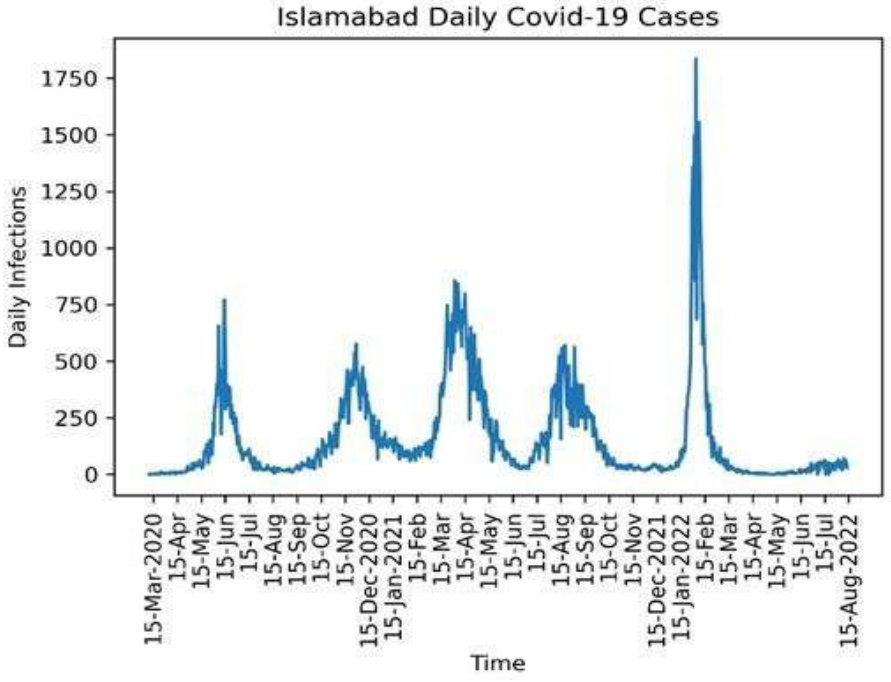


FIGURE 6.8: Visualization of daily COVID-19 cases in Islamabad.

6.3 Results

6.3.1 ARIMA Model Fit for all over Pakistan

The study first trained the ARIMA model on the available time series data. A 95 percent confidence accuracy level was utilized. To determine the prevalence of COVID-19, multiple ARIMA models were used depending on the region. The Augmented Dickey-Fuller unit test, along with ACF and PACF correlogram, showed that the prevalence of daily infection and forecasted values are not affected by seasonality. The time series data were visualized and observed for the overall data for each region. The summary of the ARIMA models used is given in Table 6.2.

TABLE 6.2: Summary of ARIMA Models used on time series data.

Region	**ARIMA (p, d, q)**
Pakistan	**(14,1,2)**
Punjab	**(14,1,1)**
Sindh	**(14,1,1)**
Khyber Pakhtunkhwa (KPK)	**(10,1,1)**
Baluchistan	**(8,1,2)**
Gilgit Baltistan (GB))	**(11,1,1)**
Azad Jammu and Kashmir (AJK)	**(14,1,1)**
Islamabad (ICT)	**(6,1,1))**

The modelled value for all regions was close to the actual ones for each region. The blue lines, which are the forecasted ones, are closer to the actual value, with a 95 percent confidence level for each region. The ARIMA Fit for each region showed lower AIC values, especially for forecasted data which showed the goodness of the fit for the next year, as shown in Figures 6.9 to 6.16.

In this section, we presented the COVID-19 case models for the above-mentioned areas of Pakistan made by the ARIMA model. The blue lines in Figure 2 show predicted cases, while the orange line represents actual COVID cases. From Figures 6.8-6.16, we see the decline in the number of cases for September 2022 for all areas except for Punjab. Moreover, it also seems to have a larger number of cases across Pakistan. The forecast and the estimate obtained are influenced by the "case" definition and the modality of data collection. The error analysis for the ARIMA model is presented next section. There is a 95 percent confidence interval for all districts. Moreover, the modelled forecasted values are close to the actual cases.

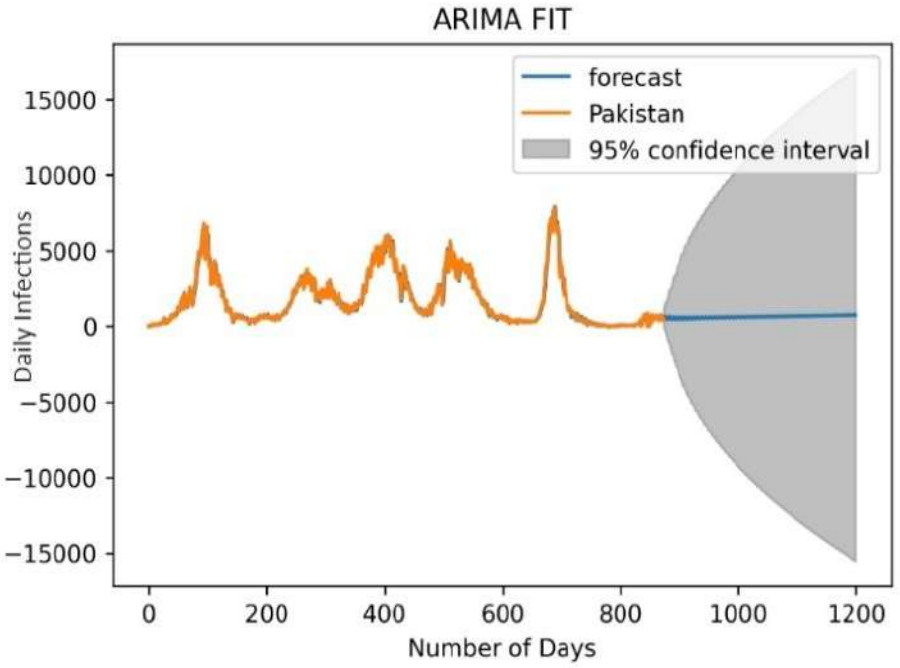


FIGURE 6.9: ARIMA FIT for 874 days of data for daily COVID-19 cases in Pakistan

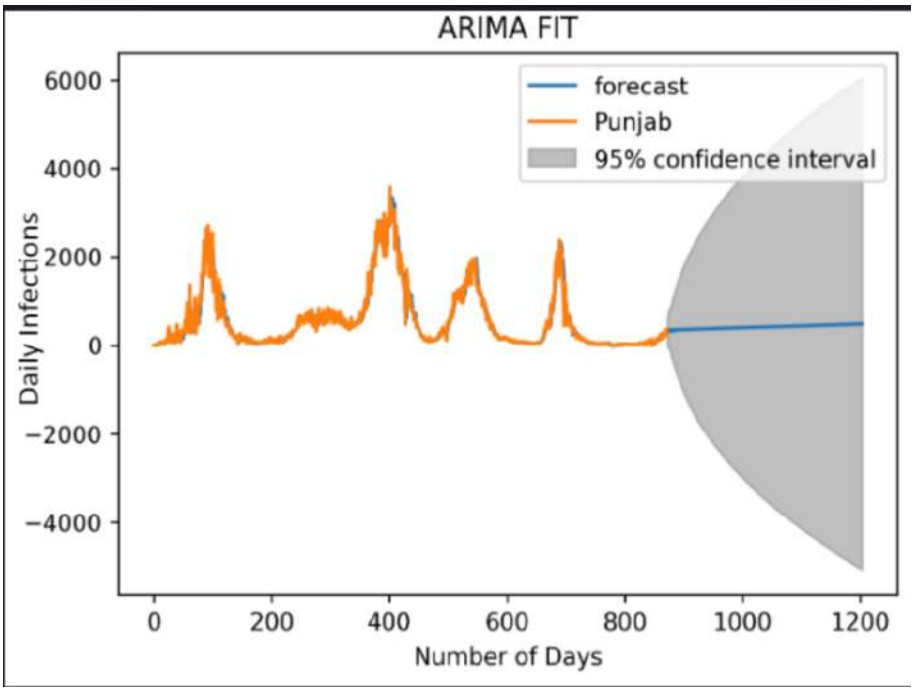


FIGURE 6.10: ARIMA FIT for 874 days of data for daily COVID-19 cases in Punjab

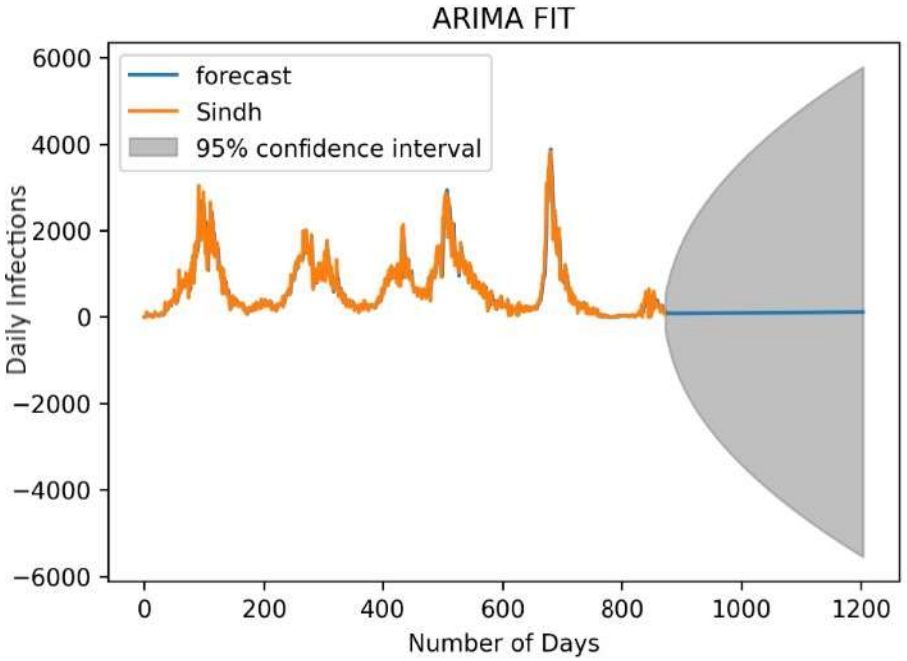


FIGURE 6.11: ARIMA FIT for 874 days of data for daily COVID-19 cases in Sindh

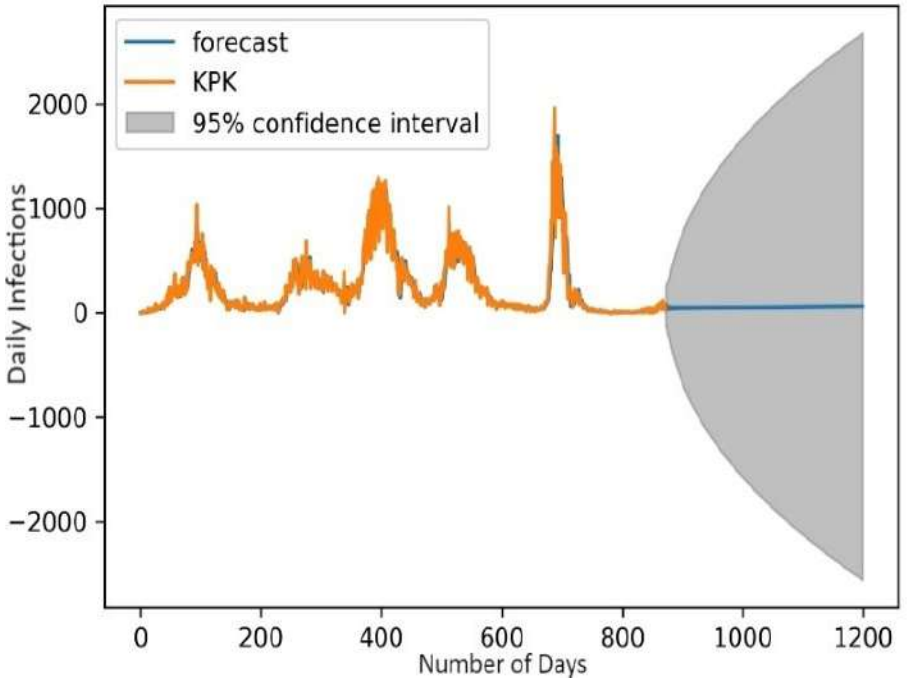


FIGURE 6.12: ARIMA FIT for 874 days of data for daily COVID-19 cases in Khyber Pakhtunkhuwa

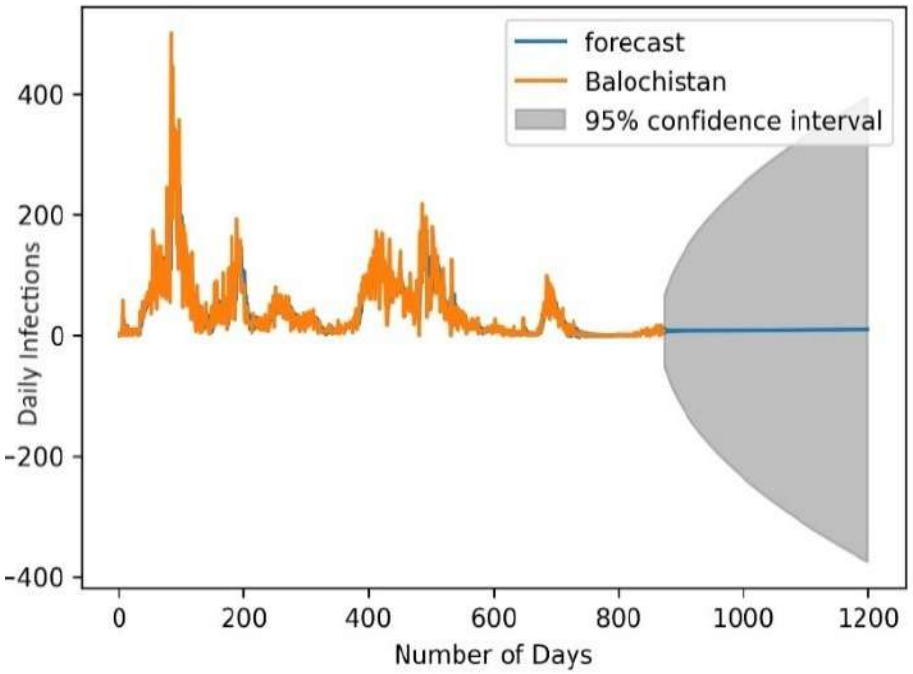


FIGURE 6.13: ARIMA FIT for 874 days of data for daily COVID-19 cases in Baluchistan

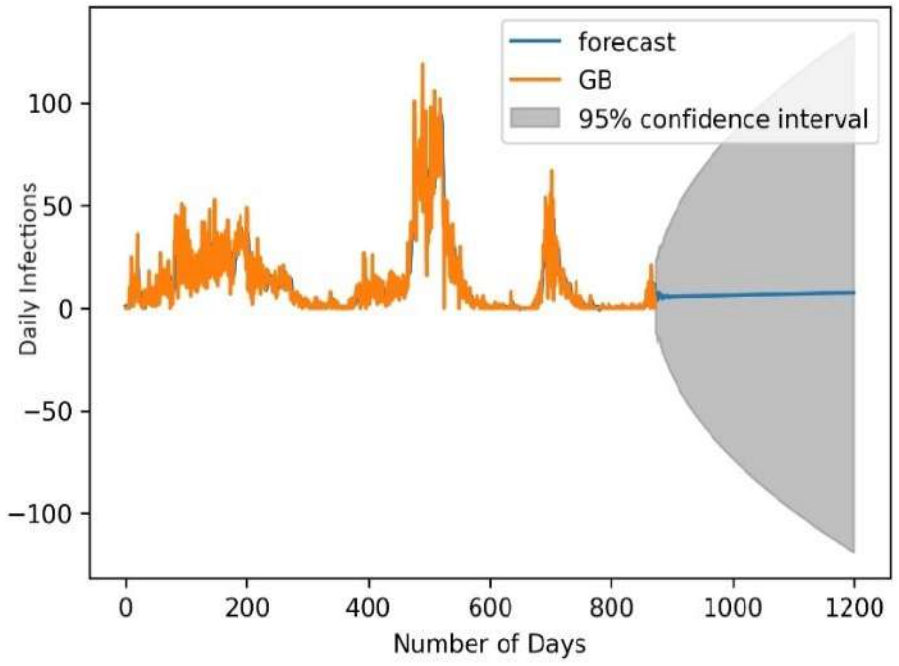


FIGURE 6.14: ARIMA FIT for 874 days of data for daily COVID-19 cases in Gilgit Baltistan.

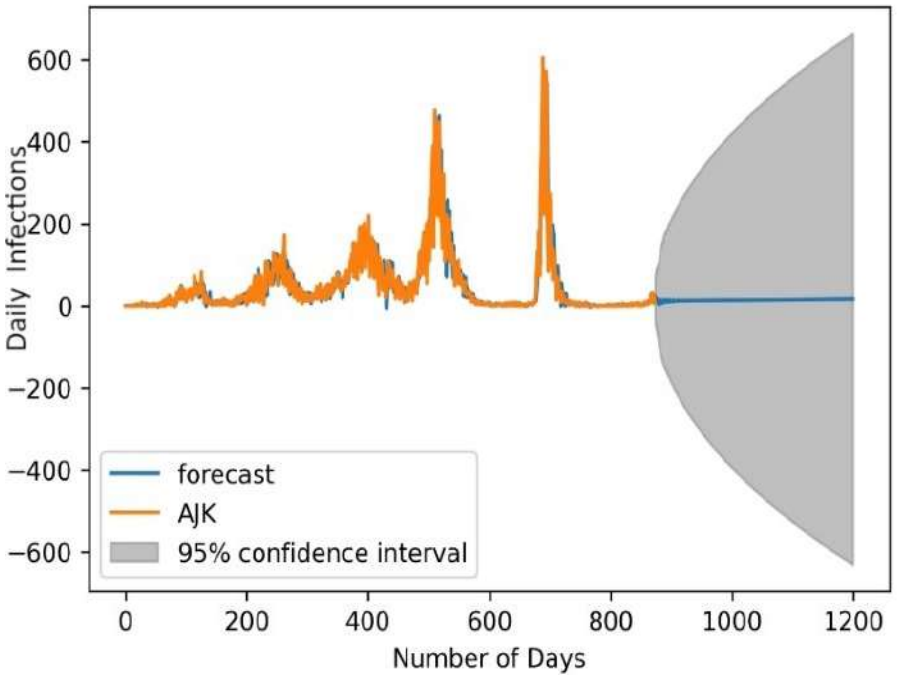


FIGURE 6.15: ARIMA FIT for 874 days of data for daily COVID-19 cases in Azad Jammu and Kashmir

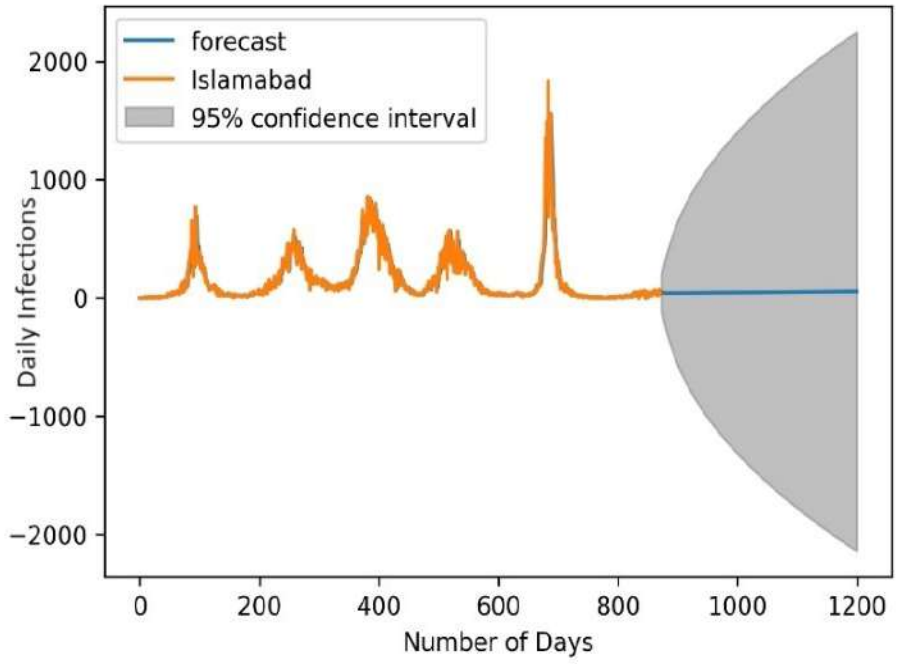


FIGURE 6.16: ARIMA FIT for 874 days of data for daily COVID-19 cases in Islamabad

6.3.2 ARIMA Error Analysis

This section presents the error analysis for the ARIMA model. The RMSE results for all regions are given in Table 6.3. The average caseload for each wave is also shown for better comparison. From the RMSE, we find the ARIMA model produces the best results for Gilgit Baltistan. Maximum error was noticed when we predicted the overall cases for the country since there were more fluctuations in the data. The ARIMA model was successful in the overall predictive modeling of the daily infection cases in all regions of Pakistan and overall Pakistan itself. The RMSE for this approach was lower for GB and Baluchistan in comparison to other regions.

Table 6.3: Error Analysis for ARIMA Model for each wave.

Area	**1**st –	**Avg -1**st	**2**nd–	**Avg - 2**nd	**3**rd	**Avg -3**rd	**4**th	**Avg - 4**th	**5**th	**Avg - 5**th
Pakistan	604	1595	510	1709	558	2977	508	2021	540	1108
AJK	24	13	24	46	37	82	35	86	56	39
Baluchistan	57	74	28	34	58	61	58	40	19	10
Sindh	438	698	365	793	306	621	302	856	243	477
Punjab	457	516	273	436	392	1385	250	596	256	280
ICT	71	84	117	175	143	304	85	157	121	125
GB	10	17	16	11	13	8	19	27	12	6
KPK	141	196	135	214	112	517	135	259	152	173

Table 6.4: Error Analysis for ARIMA Model for overall wave for each region.

Area	**Overall Wave**	**Average Cases**
Pakistan	607	1773
AJK	47	50
Baluchistan	10	41
Sindh	260	675
Punjab	250	584
Islamabad	88	158
GB	15	31
KPK	139	253

6.3.3 Logistic Growth Fit

This section analyzes the logistic fit model for the five waves of COVID-19 in Pakistan. The analysis is shown in great detail in Figure 3. The logistic growth model was more effective in helping find out the predicted cases for each wave individually. On the other hand, the logistic fit was only suitable for part of data set for Pakistan alone since there is a great variation between the actual and predicted values. The data set for 874 days was large, which caused the model to produce mirror values to reduce the error between the forecasted value and the actual case number. Therefore, the data set was divided into multiple waves and analyzed, as shown in Figures 17-21.

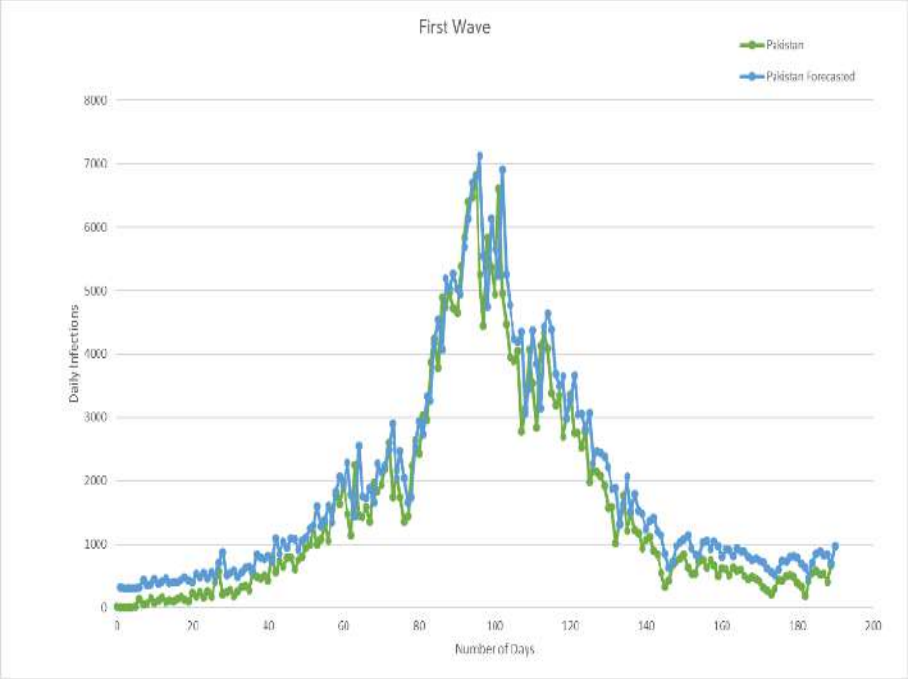


FIGURE 6.17: Logistic Growth Modeling for 1st Wave

However, even with the division of data into multiple waves, as explained in the previous section, the logistic fit was different than expected. Due to the fluctuating nature of the data, the logistic growth model fit showed a higher difference when compared to the actual data set. The logistic fit model has shown a great difference in predicting cases for each wave effectively. An overall increasing trend is visible in all five waves. However, the last three waves were not predicted as accurately as expected, as shown in Figure 6.17-21.

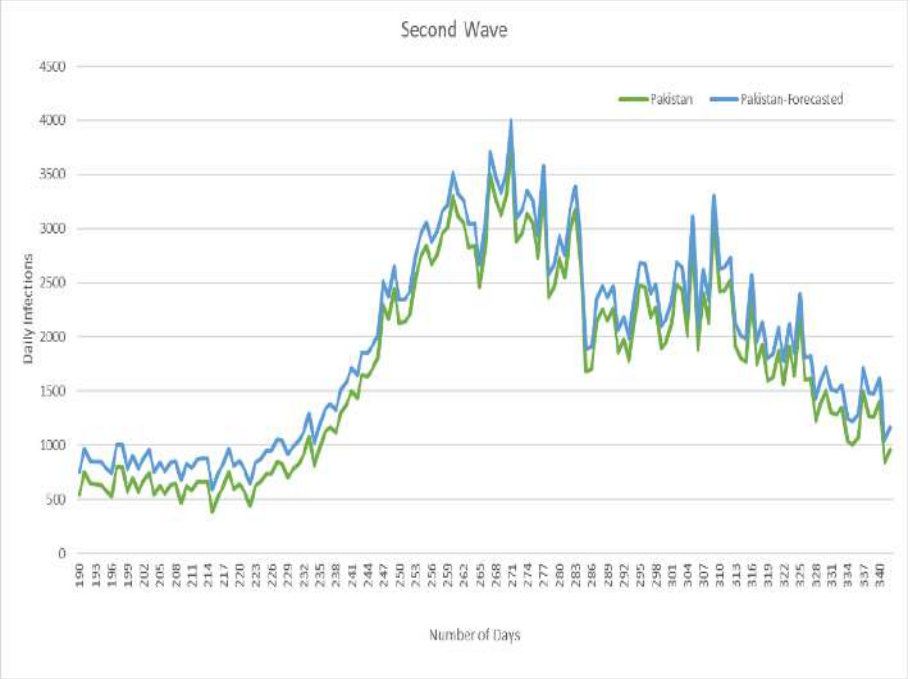


FIGURE 6.18: Logistic Growth Modeling for 2nd Wave.

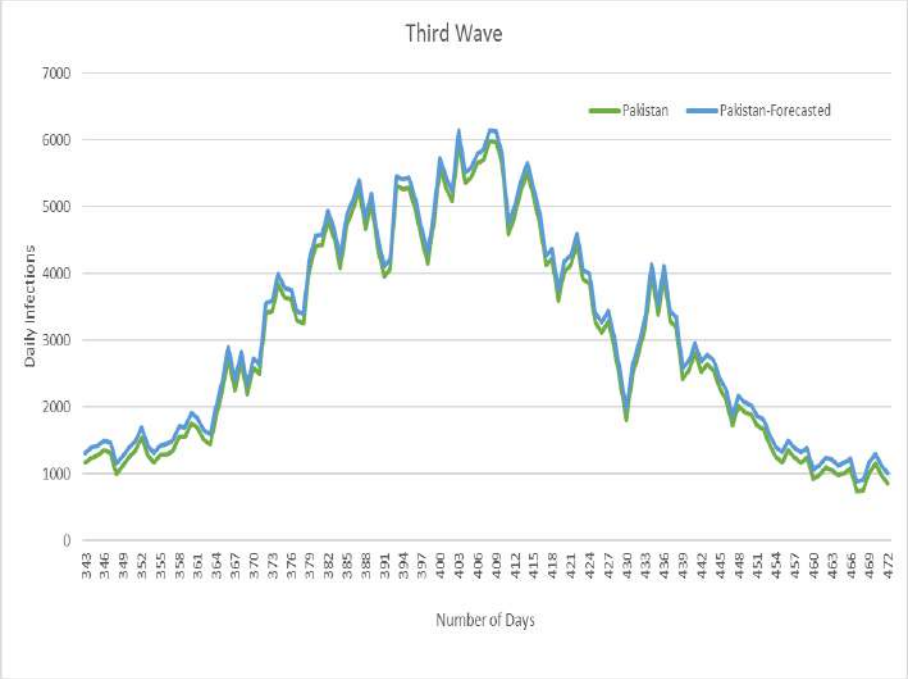


FIGURE 6.19: Logistic Growth Modeling for 3rd Wave.

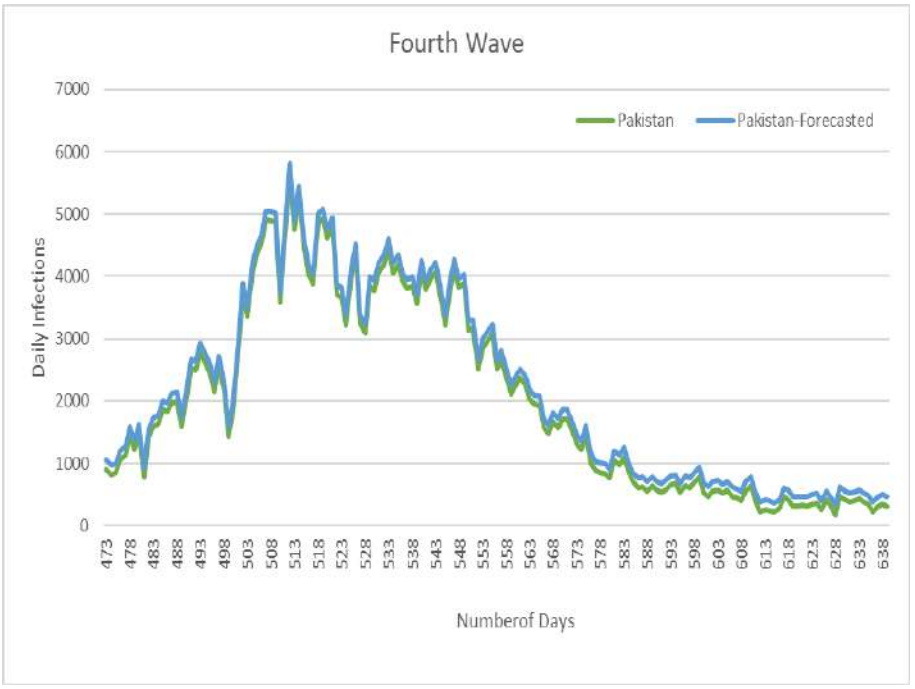


FIGURE 6.20: Logistic Growth Modeling for 4th Wave.

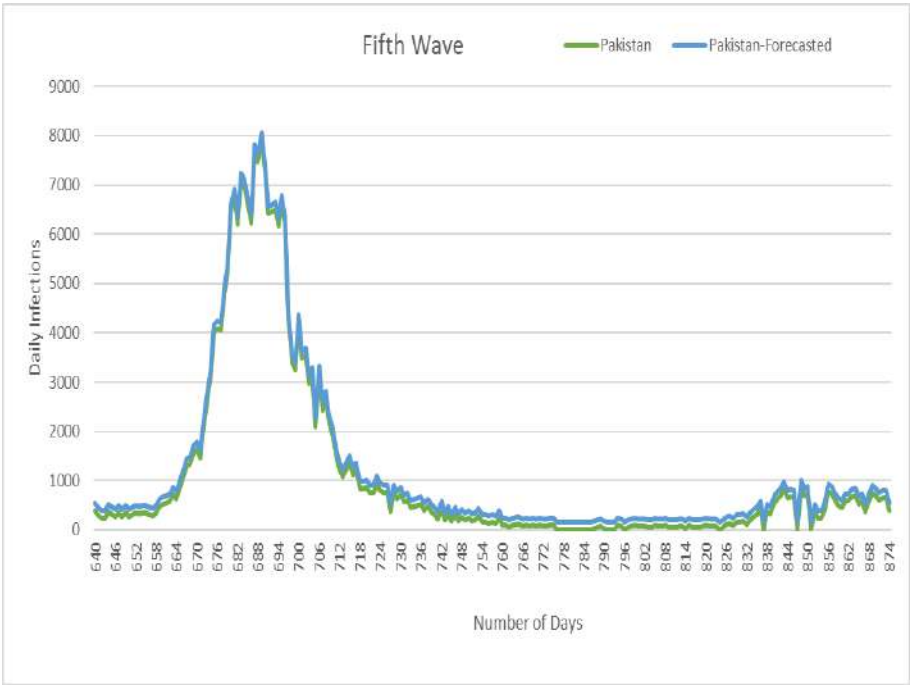


FIGURE 6.21: Logistic Growth Modeling for 5th Wave.

6.3.4 Logistic Error Analysis

This section presents the error analysis for the Logistic Prediction model. The RMSE for all waves is given in Table 4. The average daily caseload has also been given to ensure better comparisons. From the RMSE, we find the Logistic Predictions model produces the best results for the first and third waves. This approach, according to the error plot, is also be suitable for approaches where there are fewer fluctuations in the data. The maximum error was noticed when we predicted the overall cases for the country; therefore, the approach might be disregarded for the COVID-19 situation in Pakistan if the total number of cases for all waves is considered. However, for individual waves the error analysis showed better results compared to ARIMA Model. For the overall modeling for Pakistan, the logistic growth model produced inaccurate results to predict the number of daily COVID-19 cases effectively for the entire pandemic. The model tries to zero out the error. However, breaking the overall daily infection cases into smaller waves showed better results for this model compared to the overall approach. Therefore, another model would be a better option for this time-series data for modelling data for the overall pandemic.

Table 6.5: Error Analysis for Logistic Growth Model for each wave.

COVID-19 Waves	**RMSE Results**	**Average Case Load**
First Wave: 10th March 2020 to 15th September 2020	521	1595
Second Wave: 16th September 2020 to 15th February 2021	616	1709
Third Wave: 16th September 2020 to 15th February 2021	502	2977
Fourth Wave: 26th June 2021 to 9th December 2021	532	2021
Fifth Wave: 10th December 2021 to 1st August 2022 onwards	541	1108

6.3.5 Chapter Summary

This chapter has thoroughly explained the results that was obtained by using data mining as well as Machine Learning models. Overall, one could have seen that there is an improvement in the precision of estimating with the help of Machine Learning models. Logistic Growth Model showed promising results for the fluctuating data trends of the pandemic in Pakistan In the next chapter, the conclusion and future works on this research have been elaborated on.

Chapter 7

Conclusion and Future Work

The primary objective of this chapter is to propose the main findings of this research as well as the research questions. This chapter of the research paper gives a short overview regarding why is this propose research significant and future aspects of the research study.

7.1 Conclusion

For this research work, time series machine learning models are utilized for predictive modeling. Therefore, in this research study, we used the Autoregressive Integrated Moving Average and Logistic growth model for the modeling of Coronavirus cases based on the time series data accessible for the five waves for 874 days in Pakistan and its provinces. The data used starts from March 2021 to August 2022. The Autoregressive Integrated Moving Average Model demonstrated good outcomes at a 95 percent confidence interval, while the Logistic growth model demonstrated a higher precision with lower Root Mean Square Error. For the overall forecasts for Pakistan, the logistic model was not able to model predicted daily Coronavirus case numbers efficiently. The model attempts to decrease the error, producing a plot that was not accurate. Breaking the overall everyday infection cases into smaller waves displayed better outcomes for this model, in comparison to, the overall method. Nonetheless, the errors were a lower than the Autoregressive Integrated Moving Average model. This research could be substantial when it comes to managing future infectious illnesses and modeling their mitigation techniques identical to those utilized in Coronavirus, particularly for densely populated Asian nations, such as Pakistan.

7.2 Future work

In future, it is suggested to conduct a research on various Machine Learning approaches, like Deep Learning and Naive Bayes techniques to examine the trends in the Coronavirus waves further for Pakistan and different infectious diseases too. Moreover, this work will utilize an extensive data set for cross region detection and comparison. The comparison will be done with neighbouring countries like India, Afghanistan and China. Furthermore, it can further be extended to countries of similar population density as Pakistan.

Chapter 8

References

References

[1] M. K. Ali, " Forecasting COVID-19 in Pakistan." *Plos one*, Nov. 2020, vol. 15, no. 11, pp. 1-13.

[2] I, Ahmad, " Predictions of coronavirus COVID-19 distinct cases in Pakistan through an artificial neural network," Epidemiology & Infection, 2020, vol. 145, no. 1, pp. 1-17.

[3] S. F. K. Ardabil, "Covid-19 outbreak prediction with machine learning." *Algorithms,* 2020, vol. 13, no. 10, pp. 249.

[4] T. Saba, I, Abunadi, M.N. Shahzad, and A.R, Khan, Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types." Microscopy Research and Technique, 2021, vol. 57, pp. 1-13.

[5] P. Wang, X. Zheng, J. Li, and B. Zhu, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics." *Chaos, Solitons and Fractals*, 2020, vol. 139, pp. 110058.

[6] A.F. Morais, "Logistic approximations used to describe new outbreaks in the 2020 COVID-19 pandemic." *arXiv: Populations and Evolution*, 2020, pp. 1-19.

[7] F. Rojas, O. Valenzuela, and I. Rojas, "Estimation of COVID-19 dynamics in the different states of the United States using Time-Series Clustering." *edRxiv,* 2020, pp. 1-18.

[8] D.G. Chen, X. Chen, and J. K. Chen, "Reconstructing and forecasting the COVID-19 epidemic in the United States using a 5-parameter logistic growth model." *Global Health Research Policy,* 2020, vol. 5, no. 1. DOI: 10.1186/s41256-020-00152-5.

[9] C.Y. Shen, "Logistic growth modeling of COVID-19 proliferation in China and its international implications." *International Journal of Infectious Diseases*, 2020, vol. 96, pp. 582–589. DOI: 10.1016/j.ijid.2020.04.085.

[10] V. Tulshyan, D. Sharma, and M. Mittal, "An eye on the future of COVID-19: prediction of likely positive cases and fatality in India over a 30-day horizon using the Prophet model." *Disaster Medicine and Public Health Preparedness*, 2020, pp. 1–7.

[11] P. Mishra, A. Mohammad, G. Al Khatib et al., "Modelling and forecasting of COVID-19 in India." *Journal of Infectious Diseases and Epidemiology*, 2020, vol. 6, no. 5.

[12] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, and M. Ciccozzi, "Application of the ARIMA model on the COVID-2019 epidemic dataset." *Data in Brief*, 2020, vol. 29, p. 105340.

[13] F. Saif, "COVID-19 Pandemic in Pakistan: Stages and Recommendations." *medRxiv*, 2020, pp. 1-12.

[14] D. Tatrai and Z. Várallyay, “COVID-19 epidemic outcome predictions based on logistic fitting and estimation of its reliability.” *Arxiv*, 2020, pp. 1–536.
[15] S.A. Rida Ahmed, “Real-time forecast of final outbreak size of novel coronavirus (COVID-19) in Pakistan: A data-driven analysis.” *SSRN*, 2020, pp. 1-8.
[16] F. Khan, A. Saeed, and S. Ali, “Modeling and Forecasting of New Cases, Deaths and Recover Cases of COVID-19 by using Vector Autoregressive Model in Pakistan.” *Chaos, Solitons & Fractals*, 2020, vol. 140, p. 110189. DOI: 10.1016/j.chaos.2020.110189.
[17] M. Aslam, “Using the Kalman filter with Arima for the COVID-19 pandemic dataset of Pakistan.” *Data in Brief.*, 2020, vol. 31, p. 105854. DOI: 10.1016/j.dib.2020.105854.
[18] T. M. Awan and F. Aslam, “Prediction of daily COVID-19 cases in European countries using automatic ARIMA model.” *Journal of Public Health Research*, 2020, vol. 9, no. 3, pp. 227–233. DOI: 10.4081/jphr.2020.1765.
[19] M. Yousaf, S. Zahir, M. Riaz, S. M. Hussain, and K. Shah, “Statistical analysis of forecasting COVID-19 for the upcoming month in Pakistan.” *Chaos, Solitons and Fractals*, 2020, vol. 138, p. 109926. DOI: 10.1016/j.chaos.2020.109926
[20] B. Z. Peipei Wang, Xinqi Zheng, and Jiayang Li, “Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics.” *Chaos, Solitons & Fractals*, 2020, pp. 1-7.
[21] L.K. Li, X, Zhang, and Zhang B, “A comparative time series analysis and modeling of aerosols in the contiguous United States and China.” *Science of The Total Environment*, 2019, vol. *690*, pp. 799-811.
[22] H. Tandon, P. Ranjan, T. Chakraborty, and V. Suhag, “Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future.” arXiv, January 2020, pp. 1–11.
[23] G. Perone, “An ARIMA Model to Forecast the Spread of COVID-2019 Epidemic in Italy.” *SSRN Electron. J.*, 2020, DOI: 10.2139/ssrn.3564865.
[24] R. Haberman, "Mathematical models: mechanical vibrations, population dynamics, and traffic flow.” *Society for Industrial and Applied Mathematics*, 1998, pp. 1-14.
[25] G.D. Barmparis, " Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach.” *Chaos, Solitons & Fractals*, 2020, vol. 5, no. 1, pp 1-5.
[26] M.O. Daniyal, "Predictive modeling of COVID-19 death cases in Pakistan.” *Infectious Disease Modelling*, 2020, vol. 5, pp 897-904.
[27] A.S. Khakharia, "Outbreak prediction of COVID-19 for dense and populated countries using machine learning.” *Annals of Data Science*, 2021, vol. 8, pp. 1-19.
[28] S.T. Shah, "Predicting Covid-19 Infections Prevalence using Linear Regression Tool.” *Journal of Experimental Biology and Agricultural Sciences*, 2020, vol. 8, pp. 12-19.

[29] A. Ahmad, S. Garhwal, S.K. Ray, G. Kumar, S.J. Malebary, and O.M. Barukab, "The number of confirmed cases of covid-19 by using machine learning: Methods and challenges." *Archives of Computational Methods in Engineering*, 2020, vol 4, no. 3, pp 1-9.
[30] F. Rustam, A.A, Reshi, A, Mehmood, A. Ullah, S. On, B.W. Aslam, and G.S. Choi, "COVID-19 future forecasting using supervised machine learning models." IEEE *access*, 2020, vol. 3, no. 8, pp. 101489-101499.
[31] SM. Mubeen, S. Kamal, and F. Balkhi, "Knowledge and awareness regarding spread and prevention of COVID-19 among the young adults of Karachi." *J Pakistan Med Assoc.*, 2020, vol. 70, pp. 169-174.
[32] K. Abid, YA. Bari, M. Younas, S. Tahir Javaid, and A. Imran, "Progress of COVID-19 epidemic in Pakistan." *Asia Pacific J Public Health*, 2020, vol. 32, pp.154-156.
[33] Government of Pakistan. "Coronavirus in Pakistan." The government of Pakistan. https://covid.gov.pk/ (accessed Dec. 10, 2022).
[34] E. Volz, S. Mishra, M. Chand, JC. Barrett, R. Johnson, L. Geidelberg, et al, "Transmission of SARS-CoV-2 Lineage B. 1.1. 7 in England: insights from linking epidemiological and genetic data." *medRxiv*, 2021.
[35] Q. Yu, J. Liu, Y. Zhang, and J. Li, "Simulation of rice biomass accumulation by an extended logistic model including influence of meteorological factors." *Int J Biometeorol*, 2002, vol. 46, no. 4, pp. 185-191.
[36] Y.H, Hsieh, J.Y, Lee, and H.L, Chang, "SARS epidemiology modeling." *Emerg Infect Dis.*, 2004, vol. 10, no. 6, pp. 1165-1167.
[37] N. Fernandes, "Economic Effects of Coronavirus Outbreak (COVID-19) on the World Economy." *SSRN Electronic Journal*, 2020.
[38] The News. "The Global Economic Effects of Coronavirus." The News.com. https://www.thenews.com.pk/print/641469-the-global-economic-effects-of-coronavirus. (accessed Dec. 12, 2022).
[39] J. Shaman, and A. Karspeck, "Forecasting seasonal outbreaks of influenza." *Proc Natl Acad Sci USA*, 2012, vol. 109, pp. 20425-20430.

Appendices

Appendices

.1 Appendices

This section presents the code of the thesis entitled "Assessing The COVID-19 Trends In PakistanUsing Predictive Machine Learning Techniques: An Empirical Study". This code is developed by Hijab Hassan under the supervision of Dr Muhammad Asim Noor for the fulfilment of MS thesis presented to the Department of Computer Science, COMSATS University Islamabad (CUI) in Fall 2022. In this section, we have provided with the algorithms used for this study.

```
def my_logistic(t, a, b, c):
   return c /(1 + a* np.exp(-b*t) )
p0 = np.random.exponential(size=3)
p0
bounds = (0, [10000., 3., 220000])
import scipy.optimize as optim
X = np.array(df_first_wave_Pakistan['Day'] +1)
y= np.array(df_first_wave_Pakistan[area])

(a, b, c), cov = optim.curve_fit(my_logistic, X, y, bounds=bounds, p0=p0)
a , b, c
def logistic(t):
   return c /(1 + a* np.exp(-b*t) )
plt.figure(figsize=(6,4))
plt.scatter(X,y, marker='.', color='red')
plt.plot(X, logistic(X), linewidth=2)
plt.title('Logistic Model FIT on ' +area + ' Data of Covid-19')
plt.legend(['Logistic Model', 'Real Data'])
plt.xlabel('Time')
plt.ylabel('Infections')
plt.xticks(range(1,190,30),["15-Mar","15-Apr","15-May","15-Jun"], rotation=20);
z = np.array(range(215))
plt.figure(figsize=(6,4))
plt.scatter(X,y, marker='.', color='red')
plt.plot(z, logistic(z))
plt.title('Logistic Model FIT and Prediction for ' +area+ ' over Next 30 Days')
plt.legend(['Logistic Model', 'Real Data'])
plt.xlabel('Time')
plt.ylabel('Infections')

plt.xticks(range(5,250,30),["15-Mar","15-Apr","15-May","15-Jun"], rotation=20);
preds = logistic(X)
actual = df_first_wave[area]
```

FIGURE 1: LGM Algorithm

import numpy as np import matplotlib.pyplot as plt import pandas as pd	Importing required libraries
df=pd.read_excel('Location of dataset')	Loading required dataset
df.head() df.dropna(inplace=True)	Observing first 5 rows of dataset
df.columns = ['day','patients']	Removing all rows with null values
df.shape	Creating two columns
y= df.set_index('day')	Finding number of rows and columns of dataset
y.plot(figsize=(15, 6))	Setting index value to day column
plt.show()	Plotting the given dataset
from statsmodels.tsa.arima_model import ARIMA	Importing library for ARIMA
from statsmodels.graphics.tsaplots import plot_acf,plot_pacf	Importing library for acf(Autocorrelation Function) and p

FIGURE 2: ARIMA Algorithm

Figure 3: ARIMA Algorithm (Continued)

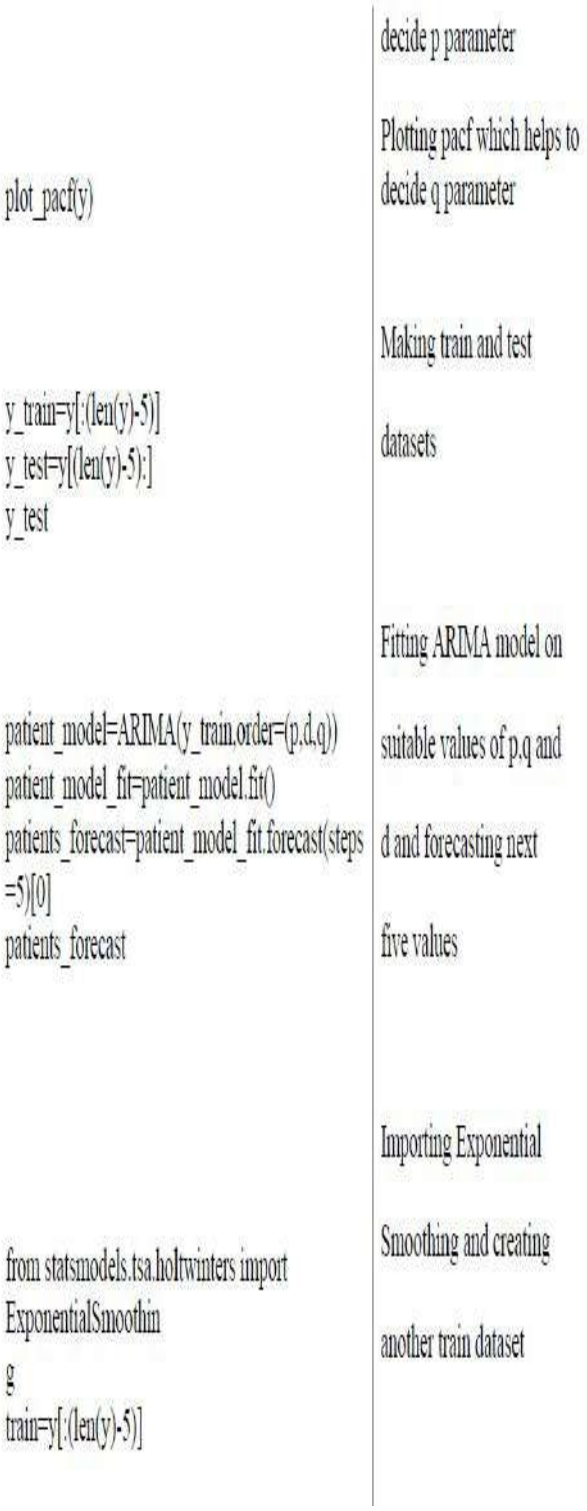


Figure 4: ARIMA Algorithm (Continued)

FIGURE 5: Data Pre-Processing

COMSATS University Islamabad

Registrar Secretariat, Academic Unit (PS)

***

No. CUI-Reg/Notif- 1909 /21/ 1410 May 25, 2021

Notification

Dean Faculty of Information Sciences and Technology has approved the supervisor and Synopsis of the student as per details given below: -

Registration No:	CIIT/FA19-RSE-018/ISB	**Name:**	Hijab Hassan
Father Name:	Hassan Akhtar	**Program:**	Master of Science in Software Engineering
Campus:	Islamabad	**Department:**	Computer Science
Supervisor:	Dr. Muhammad Asim Noor		
Title of Thesis:	Assessing the COVID-19 Trends in Pakistan using Predictive Machine Learning Techniques: An Empirical Study		

Muhammad Hanif
Deputy Registrar

Distribution:

1. Dean, Faculty of Information Sciences and Technology
2. Incharge Academics, Islamabad Campus
3. Chairman, Department of Computer Science
4. Head Department of Computer Science, Islamabad Campus
5. Controller of Examinations
6. Deputy Registrar, Academics, Islamabad Campus
7. Deputy Controller of Examinations, Islamabad Campus
8. Secretary BASAR

cc :

1. PS to Rector
2. PS to Registrar

Document Viewer

Turnitin Originality Report

Processed on: 14-Dec-2022 9:54 PM PST
ID: 1981788408
Word Count: 9981
Submitted: 1

Report By Hijab Hassan,

Similarity Index	Similarity by Source	
12%	Internet Sources:	8%
	Publications:	9%
	Student Papers:	2%

include quoted | include bibliography | excluding matches < 3 words | mode: quickview (classic) report | print | refresh | download

1% match ()
Peipei Wang, Xinqi Zheng, Jiayang Li, Bangren Zhu, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics", Chaos, Solitons, and Fractals

1% match ()
Kashif Kamran, Abid Ali, "Challenges and Strategies for Pakistan in the Third Wave of COVID-19: A Mini Review", Frontiers in Public Health

1% match (S M Nazmuz Sakib. "Software Effort Estimation for Improved Decision Making", Cambridge University Press (CUP), 2022)
S M Nazmuz Sakib, "Software Effort Estimation for Improved Decision Making", Cambridge University Press (CUP), 2022

1% match (Tahir Mumtaz Awan, Faheem Aslam. "Prediction of Daily Covid-19 Cases in European Countries Using Automatic Arima Model", Journal of Public Health Research, 2020)
Tahir Mumtaz Awan, Faheem Aslam. "Prediction of Daily Covid-19 Cases in European Countries Using Automatic Arima Model", Journal of Public Health Research, 2020

1% match ()
http://www.caa.co.uk

1% match (student papers from 27-Dec-2021)
Submitted to Higher Education Commission Pakistan on 2021-12-27

<1% match (Internet from 29-Nov-2020)
https://daten-quadrat.de/index.php?lng=de&orgN=1003

<1% match (Internet from 24-Sep-2020)
https://www.tandfonline.com/doi/full/10.1080/03670244.2015.1052426

<1% match (Internet from 17-Sep-2021)
https://www.tandfonline.com/doi/full/10.1080/02626667.2012.714468

<1% match (Internet from 07-Sep-2017)
http://openaccess.city.ac.uk

<1% match (Shah Faisal, Junaidi Khotib, Elida Zairina. "Knowledge, attitudes, and practices (KAP) towards COVID-19 among university students in Pakistan: a cross-sectional study", Journal of Basic and Clinical Physiology and Pharmacology, 2021)
Shah Faisal, Junaidi Khotib, Elida Zairina. "Knowledge, attitudes, and practices (KAP) towards COVID-19 among university students in Pakistan: a cross-sectional study", Journal of Basic and Clinical Physiology and Pharmacology, 2021

<1% match (Internet from 09-Oct-2022)
https://ojs.piscomed.com/index.php/FF/issue/download/152/20

<1% match (Internet from 14-Dec-2022)
https://www.frontiersin.org/articles/10.3389/fpubh.2022.922795/full

<1% match (Musa Khan, Gul Muhammad Khan. "COVID-19 Spread Prediction and Its Impact on the Stock market price", 2022 2nd International Conference on Artificial Intelligence (ICAI), 2022)
Musa Khan, Gul Muhammad Khan. "COVID-19 Spread Prediction and Its Impact on the Stock market price", 2022 2nd International Conference on Artificial Intelligence (ICAI), 2022

<1% match (Internet from 28-Nov-2022)
https://www.scilit.net/articles/search?q=reference_ids%3A%2812399715 7%29&sort=Newest

<1% match (Internet from 25-Oct-2022)
http://ir.knust.edu.gh

<1% match (student papers from 31-Jan-2013)
Submitted to Universiti Teknologi Malaysia on 2013-01-31