# Future Forecasting Analysis on COVID-19


| | |
|---|---|
| **Sanjida Akter Ishita** | **170204089** |
| **Md Ashfaqul Azam Chowdhury** | **170204097** |
| **Fariha Zaman** | **170204098** |


**Project Report**

**Course ID: CSE 4214**

**Course Name: Pattern Recognition Lab**

**Semester: Spring 2021**



**Department of Computer Science and Engineering**

**Ahsanullah University of Science and Technology**

Dhaka, Bangladesh


March 2022

# Future Forecasting Analysis on COVID-19

Submitted by


| | |
|---|---|
| **Sanjida Akter Ishita** | **170204089** |
| **Md Ashfaqul Azam Chowdhury** | **170204097** |
| **Fariha Zaman** | **170204098** |


Submitted To


**Faisal Muhammad Shah**, Associate Professor

**Md. Tanvir Rouf Shawon**, Lecturer

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology




**Department of Computer Science and Engineering**

**Ahsanullah University of Science and Technology**

Dhaka, Bangladesh


March 2022

# ABSTRACT

COVID-19 is a pandemic that has affected over 170 countries around the world. The number of infected and deceased patients has been increasing at an alarming rate in almost all the affected nations. Forecasting techniques can be inculcated thereby assisting in designing better strategies and in taking productive decisions. These predictions might help to prepare against possible threats and consequences. The term Forecasting about COVID-19 means the estimation of affected people, dead people, recovered people etc in different countries over the world. Since the situation is going out of control, it is better to get in knowledge about the cases around our cities or counties for stopping spread. The forecasting techniques cannot stop to transmission this virus absolutely but helps to predict cases around us so that people get aware. In this way, transmission of this virus can become under-controlled and the situation can be handled. We used the Novel Corona Virus 2019 Dataset [1] for our project, collected from the Kaggle. The dataset contains information about countries and their corresponding nine unique attributes. We have evaluated linear regression, decision tree regression, K-nearest neighbor regression, random forest regression and the results show that decision tree regression algorithm performred well on the dataset.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

A little over a year after the official announcement from the WHO, the COVID-19 pandemic has led to dramatic consequences globally. Today, millions of doses of vaccines have already been administered in several countries. However, the positive effect of these vaccines will probably be seen later than expected.The primary goal of the project is to provide the World Health Organisation (WHO) with an early prediction tool for the dissemination of new corona viruses. Most people who have coronavirus disease 2019 (COVID-19) recover completely within a few weeks. But some people even those who had mild versions of the disease continue to experience symptoms after their initial recovery. Many scientists expect the coronavirus will stick around indefinitely, but they also believe the illness it causes will become far less serious. But there is a fact that people should be more careful and spread awareness all over. Death rate, affected rate can make people more aware and make the observation of future. In this project, we worked on the dataset to predict Covid-19 cases all over the world. This project is about analysis on further transmission of Covid-19 or coronavirus and that could be useful in stopping its further spread. [1]

# Chapter 2

# Literature Reviews

Several scholars used the machine learning (ML) method to future forecasting analysis using Novel Corona Virus 19 dataset. [1] The Novel Corona Virus 19 dataset having: 8 attributes,229 records describing countries. Some closely related works are discussed in this section.

## 2.1 COVID-19 Trend Analysis using Machine Learning Techniques

### Author

Abhishek Jaglan
Daksh Trehan
Priyansh Singhal

### Approach

Here, Simple Linear Regression, Polynomial Linear Regression is used. The accuracies obtained for different cases using polynomial linear regression. The model trained and tested using polynomial linear regression is used to predict the total number of cases, death cases, recovered cases on each day. The dataset that has been utilized in prediction is fetched from data archive for “2019 Novel CoronaVirus Visual Dashboard”. The dataset conscripted

is continuous dataset and therefore, is well suited for regression analysis as it needs to predict from continuous dependent variables from various independent ones. The relation between dependent and independent variables can be defined by coefficient of both variables in regression mathematical statement.
The research paper successfully shows the analysis of the gathered covid-19 data of world. Predicting further transmission of Covid-19 or coronavirus could be useful in stopping its further spread. The 6 paper purposefully displays the comprehensive steps taken to implement the data dashboard for better understanding of the data and interactive information exchange which will be helpful in further taking necessary steps to manage the resources for its containment. [2]

## 2.2 Forecast and prediction of COVID-19 using machine learning

[3]

**Author**

Deepak Painuli
Divya Mishra
Suyash Bhardwaj
Mayank Aggarwal

**Approach**

ARIMA, COVID-19, Extra tree classifier, Forecast, Machine learning, Prediction, Random forest classifier, Time series is used in this research paper. Here used the most widely used forecasting method, called the ARIMA model for time series forecasting. ARIMA is used for time series data to predict future trends.Forecasts depend completely on past trends, so forecast values can-

not be guaranteed. Various ML techniques are used to predict and forecast future events. Some ML techniques used for prediction are support vector machine, linear regression, logistic regression, naive Bayes, decision trees (random forest and ETC), K-nearest neighbor, and neural networks (multi-layer perceptron)

**Evaluation**

| ARIMA → | Parameter (p, d, q) configuration | | |
|---|---|---|---|
| State ↓ | 5,1,0 | 3,1,0 | 1,1,0 |
| Maharashtra | 320.31 | 284.49 | 128.82 |
| Gujarat | 454.54 | 394.30 | 379.00 |
| Delhi | 287.27 | 319.62 | 125.04 |
| Rajasthan | 48.29 | 75.30 | 78.25 |
| Madhya Pradesh | 92.33 | 107.07 | 116.64 |
| Tamil Nadu | 105.02 | 98.24 | 51.59 |
| Uttar Pradesh | 154.22 | 37.23 | 144.01 |
| Telangana | 12.66 | 17.97 | 99.49 |
| Andhra Pradesh | 13.59 | 14.28 | 29.93 |
| West Bengal | 10.90 | 3.38 | 21.28 |
| All India | 1299.87 | 758.60 | 253.60 |
| | Error rate (root mean square error value) ↑ | | |

Figure 2.1: Error rate (root mean square error [RMSE])

The error rate (root mean square error [RMSE]) of the model for different states is shown using the ARIMA model. Entries in bold show the lowest RMSE for the state of a particular pdq configuration. The lowest value of RMSE is treated as the best configuration of ARIMA to forecast future values for a particular state. According to the values given above, they have selected two states, Telangana and West Bengal, and the total cases for India.

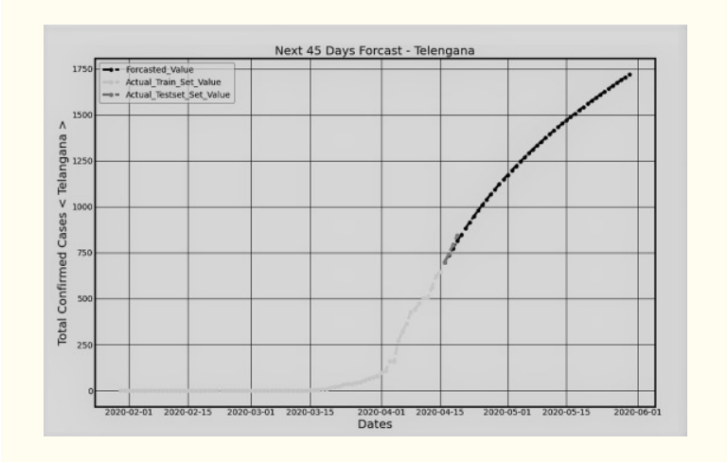


Figure 2.2: Forecast for Telangana.

Here shows the forecast for Telangana state based on data up to Apr. 19, 2020. Dark gray dots signify the actual training set (past real observation) upon which the model was trained and green (gray in printed version) signifies the actual testing set (partial data points from the dataset) for which the furcated value was validated (see overlapping area of light gray dots); pink (black in printed version) is the future forecast. [3]

# Chapter 3

# Data Collection & Processing

For our project we used a dataset named 'Novel Corona Virus 2019 Dataset' which we have collected from Kaggle. [1] From World Health Organization - On 31 December 2019, WHO was alerted to several cases of pneumonia in Wuhan City, Hubei Province of China. The virus did not match any other known virus. This raised concern because when a virus is new, we do not know how it affects people.So daily level information on the affected people can give some interesting insights when it is made available to the broader data science community. Johns Hopkins University has made an excellent dashboard using the affected cases data. Data is extracted from the google sheets associated and made available here.Latest number of affected cases In this dataset there have 8 columns and 306429 rows. There has 8 unique attributes and 229 countries. 8 unique attributes contain SNo, ObservationDate, Province/State, Country/Region, Last Update, Confirmed, Deaths, Recovered, object. Description of these attributes are given below:

Sno - Serial number
ObservationDate - Date of the observation in MM/DD/YYYY
Province/State - Province or state of the observation (Could be empty when missing)
Country/Region - Country of observation
Last Update - Time in UTC at which the row is updated for the given province

or country. (Not standardised and so please clean before using it)
Confirmed - Cumulative number of confirmed cases till that date
Deaths - Cumulative number of of deaths till that date
Recovered - Cumulative number of recovered cases till that date

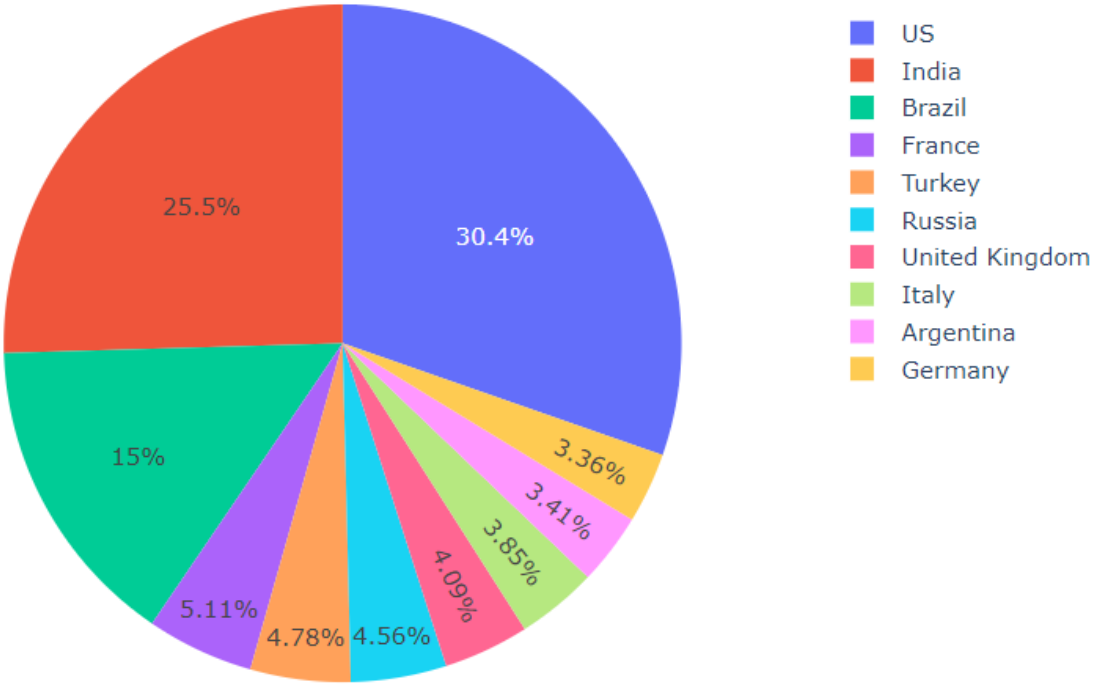


Figure 3.1: Contribution of top 10 countries so far

Here portraits the contribution of 10 countries of dataset according to covid cases. In this chart, the number of Covid cases is comparetively high. After US there is India and then Brazil.

## 3.1 Correlation using visualization

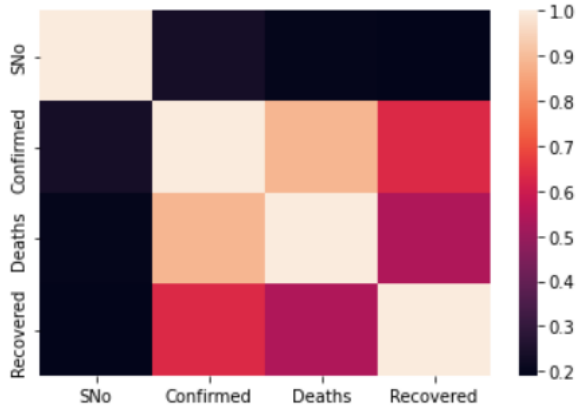


Figure 3.2: Correlation using visualization

Here shows the correlation among recovered, deaths, confirmed. If correlation is around 1.0 then it means the good result and below this shows the less correlated.

## 3.2 Similar spread comparison between countries

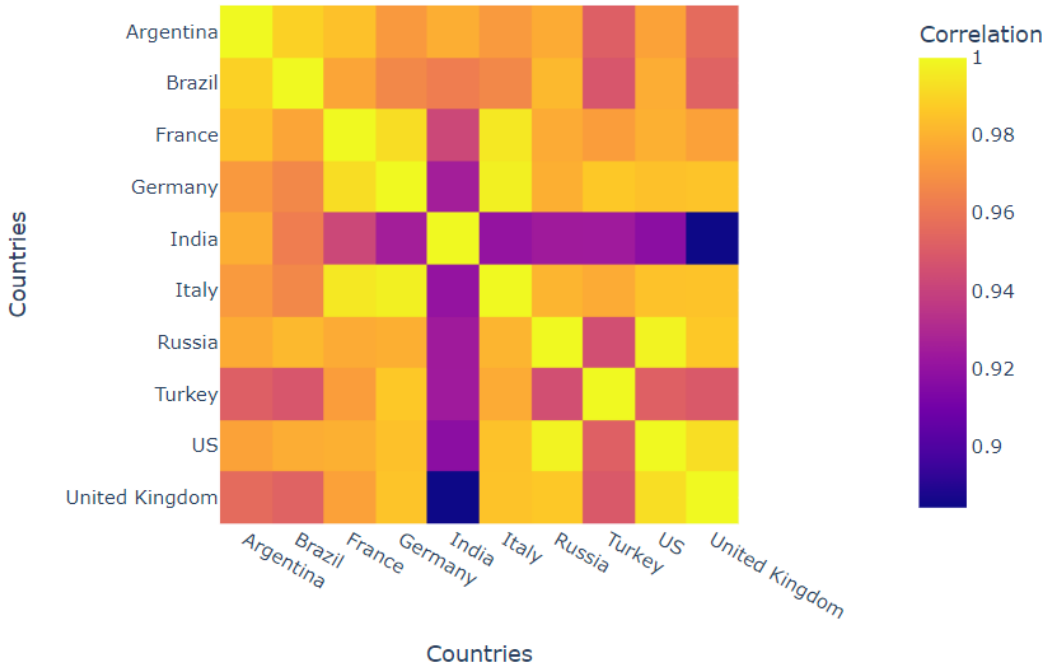


Figure 3.3: Similar spread comparison

Here shows the correlation among top 10 countries. If correlation is around 1.0 then it means the good result and below this shows the less correlated.

# Chapter 4

# Methodology

For our project we used four machine learning algorithms like Linear Regression, Decision Tree Regression, K Nearest Neighbor (KNN) Regression, Random Forest Regressor and Decision Tree (DT). For this we preprocessed the data and split the data for training and testing the model. For training we used 80% of the data and rest of the 20% for testing purpose. After training the models, we evaluated the models with testing data.

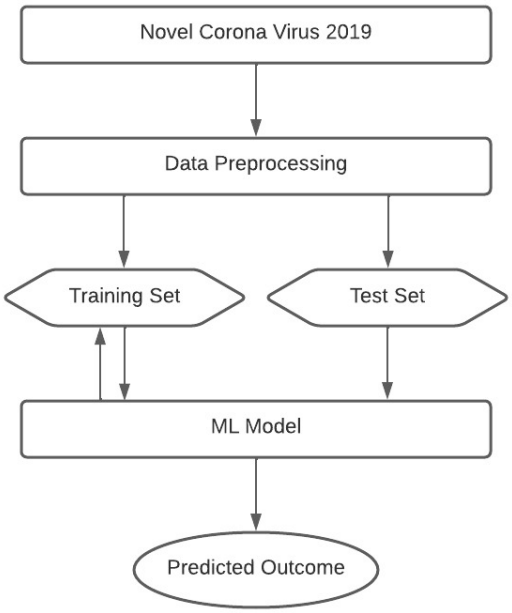


Figure 4.1: Methodology

# Chapter 5

# Experiments and Results

## 5.1 Linear Regression

The aim class concentrates on individual regression simulation characteristics. It may also be used to define and model the relationship between independent variables and dependent. The most useful computer method for mathematical analysis of the machine learn is linear regression type regression simulation. A linear regression observation relies on two values, one on the dependence and one on the isolation. Linear Regression defines a linear relation between these variables' dependency and independence. [4]

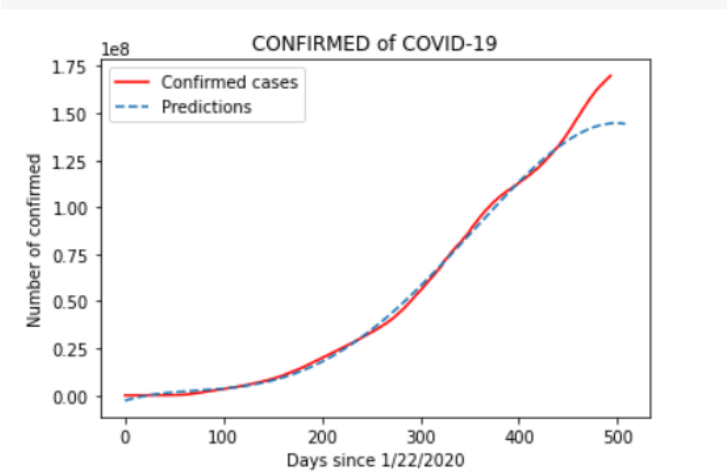


Figure 5.1: Comparison between confirmed cases and predictions

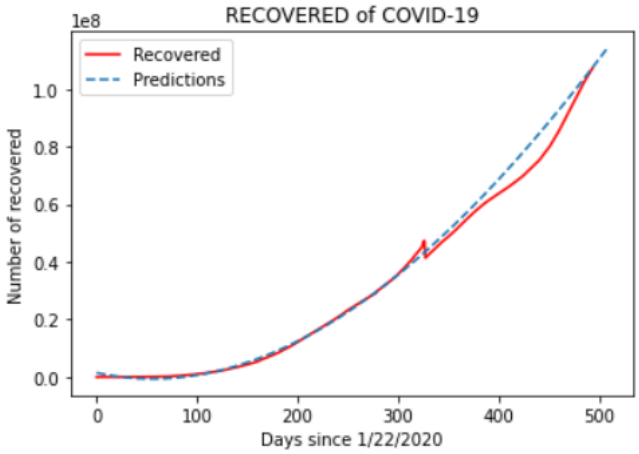


Figure 5.2: Comparison between recovered cases and predictions

| | |
|---|---|
| Mean Absolute Error | 978.6087171773605 |
| Mean Squared Error | 8723087.431598691 |
| Root Mean Squared Error | 2953.4873339153987 |

Table 5.1: Result table of Linear Regression

## 5.2 Decision Tree Regression

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

The decision trees is used to fit a sine curve with addition noisy observation. As a result, it learns local linear regressions approximating the sine curve.

| | |
|---|---|
| Mean Absolute Error | 449.2677827453796 |
| Mean Squared Error | 4232391.1501049725 |
| Root Mean Squared Error | 2057.2776064753566 |

Table 5.2: Result table of Decision Tree Regression

## 5.3 KNN Regressor

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood. Data science or applied statistics courses typically start with linear models, but in its way, K-nearest neighbors is probably the simplest widely used model conceptually. KNN models are really just technical implementations of a common intuition, that things that share similar features tend to be, well, similar. This is hardly a deep insight, yet these practical implementations can be extremely powerful, and, crucially for someone approaching an unknown dataset, can handle non-linearities without any complicated data-engineering or model set up.

| | |
|---|---|
| Mean Absolute Error | 464.57702574813163 |
| Mean Squared Error | 2682895.5762220845 |
| Root Mean Squared Error | 1637.9546929698893 |

Table 5.3: Result table of KNN Regression

## 5.4 Random Forest Regression

Random forest is one of the most popular algorithms for regression problems (i.e. predicting continuous outcomes) because of its simplicity and high accuracy. In this guide, we'll give you a gentle introduction to random forest and the reasons behind its high popularity. Random forest regression is a popular algorithm due to its many benefits in production settings: These are Extremely high accuracy, Scales well, Interpretable and Easy to use.

| | |
|---|---|
| Mean Absolute Error | 13734.30534651742 |
| Mean Squared Error | 12872964338.840065 |
| Root Mean Squared Error | 113459.08662967486 |

Table 5.4: Result table of Random Forest Regression

# Chapter 6

# Future Work and Conclusion

In this project we have used the Novel Corona Virus 2019 dataset [1] and experimented with four machine learning like Linear Regression, Knn Regression, Decision Tree Regression and Random Forest Regression. We trained and evaluated the stated model and found out Decision tree model performed better than other models on forecasting analysis from the Novel Corona Virus 2019 dataset.The Decision tree model achieved mean absolute error is 449.2677827453796 which is comparetively low then other model's mean absolute error.

In future, we plan to collect a more enriched dataset which will help us to analysis on future forecasting with higher confidence. In addition to this, we also plan to extend this work to evaluate time series forecasting to process of analyzing time series data using statistics and modeling to make predictions and inform strategic decision-making. We will also use other higher level models and deep learning techniques perform.

# References

[1] N. C. Virus, "Dataset," *Kaggle (2020). Available online at: https://www. kaggle. com/sudalairajkumar/novel-corona-virus-2019-dataset*, 2019.

[2] A. Jaglan, D. Trehan, U. Megha, and P. Singhal, "Covid-19 trend analysis using machine learning techniques," *Int J Sci Eng Res*, vol. 11, no. 12, pp. 1162–1167, 2020.

[3] D. Painuli, D. Mishra, S. Bhardwaj, and M. Aggarwal, "Forecast and prediction of covid-19 using machine learning," in *Data Science for COVID-19*, pp. 381–397, Elsevier, 2021.

[4] R. K. Mojjada, A. Yadav, A. Prabhu, and Y. Natarajan, "Machine learning models for covid-19 future forecasting," *Materials Today: Proceedings*, 2020.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Saturday 29th April, 2023 at 7:30pm.