International Journal on Emerging Technologies **11**(4): 56-61(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Predicting Rise and Spread of COVID-19 Epidemic using Time Series Forecasting Models in Machine Learning

Ch. V. Raghavendran[1], G. Naga Satish[2], Vempati Krishna[3] and Shaik Mahaboob Basha[4]
[1]Professor, Department of Information Technology,
Aditya College of Engineering & Technology, Surampalem (Andhra Pradesh), India.
[2]Associate Professor, Department of Computer Science and Engineering,
BVRIT HYDERABAD College of Engineering for Women, Hyderabad (Telangana), India.
[3]Professor, Department of Computer Science and Engineering,
TKR College of Engineering and Technology, Hyderabad (Telangana), India.
[4]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering,
Saveetha Institute of Medical and Technology Sciences, Saveetha nagar, Thandalam, Chennai, India.

(Corresponding author: Ch. V. Raghavendran)
(Received 27 May 2020, Revised 18 June 2020, Accepted 20 June 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The rapid spread of COVID-19 epidemic has attracted worldwide attentions since December, 2019. In this 21st century with most advanced technology available, we are not able to stop it. In this paper, we will study how the Machine Learning (ML) techniques are useful for analyzing and predicting the rise and spread of COVID-19. We study the rise of COVID-19cases in countries all over the world. We further aim to compare the spread of COVID-19 in selected countries and try to predict the possible cases up to date. We use the datasets available from John Hopkins University to study how accurately one can predict the rise with mathematical modeling. We use different Regression techniques, Time series forecasting techniques like Holt's model, ARIMA model to analyze on the rise and extent of the virus. The challenge in predicting the possible cases with traditional models is with high error value and is reduced by using Holt's, ARIMA models.

Keywords: COVID-19, Corona virus, Machine Learning, Linear Regression, Support Vector Machine, Time series, Holt's model, ARIMA model.

I. INTRODUCTION

At the completion of year 2019, a novel Corona VIrus Disease 2019 (COVID-19) pneumonia arisen in the Wuhan, China [1-4]. On 24th January, 2020, Huang *et al*. explained the medical features of forty one patients with COVID-19. In this they specified the collective inception of indications which are - fever, cough, myalgia, or weakness. These had "pneumonia" and their chest CT inspection indicated irregularities. The impediments involved "acute respiratory distress syndrome, acute heart injury, and secondary infections". Among these patients, 13 were ICU, and 6 were died. For the first time the Kok-KH [5] team at the "University of Hong Kong" found the confirmation of human-to-human spread of COVID-19. This virus has initiated severe community health protection complications and later converted an international concern [6-8]. The severe situation puts onward new necessities for the stoppage and regulator plan. As a saying, "*always prevention is better than cure*" and it's time to "*stay home and stay safe*".

As the COVID-19 virus eruption continues to extent through the globe, enterprises and scholars are looking to use Machine Learning (ML) as a way of speaking the experiments of the virus. Computer experts and ML scholars all over the globe have been working together and functioning widely to discovery solutions to resolve problem slinked to the coronavirus. They are working on the datasets and preparing algorithms to study from the dataset. Even though, the statistical methods and procedures are available in the literature; the genuine intention for the explosion of ML techniques is because of data, higher system power, free software and structures. From industries like manufacturing and electricity to healthcare and learning, machine learning has modernized them.

In this paper we will analyze the rise and spread of COVID-19 virus in selected countries around the world. We use regression techniques such as linear regression, support vector machines and various time series predicting models to predict the rise and spread of this virus. The paper is organized as; the next chapter will be on machine learning techniques, section III on dataset used in the analysis, section IV covers the implementation of regression techniques. In the section V we compare the results and visualize the results and concluding remarks in the section VI.

II. MACHINE LEARNING

Machine learning is a data science method that explains computers to do what comes naturally to humans and animals: learning from knowledge. These procedures use computational techniques to "learn" information straight from data without trusting on a prearranged equation as a model. With the increase in the samples, the performance of the algorithms also increases.

A. Regression

Regression algorithms are mainly used in predictive modeling. It analyzes the relationship between the target variable and the predictor variable.

Regression analysis are mainly applied on real time scenarios like for casting, prediction, time series modeling, finding the cause-effect relationship. Linear regression and logistic regression are considered to be the best algorithms among all forms of regression analysis. Other algorithms include "*Ordinary Least Squares Regression (OLSR), Stepwise Regression, Multivariate Adaptive Regression Splines (MARS) and Locally Estimated Scatter plot Smoothing (LOESS)*" [9-11].

B. Support Vector Machines (SVM)

Support Vector Machines are one of the most widely used algorithms in Machine Learning. In addition to the improved performance in many areas, Support Vector Machines have the added benefit of being simpler to analyze theoretically. Furthermore, it shows more clearly what learning is about, rather than the complicated way.

The notion behind SVM is building a splitting line, plane or hyper plane, which sets two different classifications apart. First, the convex hull for the occurrences fitting to each classification is calculated. Then the line between the points closest to the opposite hull is drawn, and the splitting plane is clear as the tangent at the median of the line.

C. Time series forecasting Models

A time series is a section of values on the identical measure indexed by a time like factor. Time Series predicting is an essential part of the knowledge and there are numerous applications in the real world. Perfect predicting is a necessary element for many administration assessments. Time series data can show a variation of arrangements, and it is frequently supportive to divide a time series into numerous components, each representing an essential pattern group.

Trend – If there is a long-standing rise or fall in the data, then it indicates the trend. This may be linear or nonlinear. This is also referred as "changing direction".

Seasonal – If the time series is exaggerated by seasonal issues like time of a year or day in a week, then this indicates a seasonal form. This is a fixed one and frequency is also identified.

Cyclic – When the data exhibit rises and falls independent of a static frequency, it indicates a cycle. This type of oscillations is generally due to economic conditions, and linked to the "business cycle".

Univariate time series – This states to a time series that involves of single (scalar) interpretations recorded consecutively over time.

Multivariate time series – This is used to model and describe the interfaces and co-movements between a set of time series variables.

Holt's linear model

Holt [12] extended simple exponential smoothing to permit the predicting of data with a trend. This technique includes a prediction equation and two smoothing equations – for the level and trend:

$\text{Forecast equation } \hat{y}_{t+h|t} = l_t + hb_t$

$\text{Level equation} l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$

$\text{Trend equation } b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1}$

where $l_t$means an "estimation of the level of the sequence" at time t, b_t means an "estimation of the trend" of the series at time t, α is the "smoothing parameter for the level", $0 \leq \alpha \leq 1$, and β^* is the "smoothing parameter for the trend", $0 \leq \beta^* \leq 1$.

Holt's Winter Model

Holt and Winters [12-13] have included "*seasonality*" as an extension to the existing Holt's method. This new seasonal technique includes the prediction equation and three smoothing equations – first equation is for level l_t, second is for the trend b_t, and the last one is seasonal element b_t, with respective "*smoothing parameters*" α, β^* and γ. In this model m is used to represent the occurrence of the seasonality, (no. of seasons/year). Based on the seasonal element, two deviations are there in this method. If the seasonal deviations are approximately constant in the series, then "Additive Method" is used. If the seasonal deviations are varying proportionate to the level of the series, then "Multiplicative method" is used.

ARIMA Model

ARIMA is an abbreviation that means "Auto-Regressive Integrated Moving Average". ARMA models are commonly used in time series modeling [15]. Specifically,

Auto Regression (AR) – This uses the needy correlation among an observation and specific number of lagged observations.

Integrated (I) – This is the use of differencing of raw observations to mark the time series fixed.

Moving Average (MA) – This uses the dependence among an observation and an outstanding error from a moving average model applied to lagged observations.

The above said three mechanisms are explicitly indicated in the model in the form of parameters. The ***AR*** and ***MA*** are two extensively used linear models that work on stationary time series, and ***I*** is a preprocessing technique to "stationarize" time series if required. ARIMA model will be constituted to accomplish the task of an ARMA model, and also AR, I, or MA models.

III. DATA SET DESCRIPTION

Regarding data related to COVID-19, there are number of official and official sites are available. "John Hopkins University's Center for Systems Science and Engineering (JHU CSSE)" dataset is the most commonly used dataset. We have used the time series and combined data for the analysis. Novel Corona Virus 2019 data is distributed across the next four files:

— time_series_19-covid-confirmed_global.csv
— time_series_19-covid-deaths_global.csv
— time_series_19-covid-recovered_global.csv
— covid_19_data.csv

The dataset consists of daily occurrence data of COVID-19 from 22/01/ 2020 to03/05/2020. The dataset has everyday case reports and everyday time series tables. The time series summary data in CSV format is considered in this paper. The data is available for confirmed, death and recovered cases of COVID-19 with six features – province/state, country/region, last update, confirmed, death and recovered cases in three files. This dataset is updated on daily basis [14].

The Fig. 1 illustrates the growth rate of COVID-19 active cases from February, 2020 to 03 May, 2020. Active cases are arrived by subtracting recovered and death cases from the confirmed cases on day to day basis. The Fig. also indicates that growth in no. of active cases is a clear indication that the recovered or death case

number is falling in comparison to confirmed cases number. The Fig. 2 presents the growth of all types of cases – Confirmed, Recovered, Death cases in all over the world. From the observed exponential growth of the virus, it is to understand that it is to be controlled. It is also clear from the figure that the recovered cases are increasing as comparing with confirmed cases. A positive sign is death cases are not increasing exponentially. The Fig. 3 presents the rise of death, recovered and confirmed cases on day by day since from February, 2020.This is calculated on discrete difference method and the results are

Every day avg. increase in confirmed cases: 34041
Every day avg. increase in recovered cases: 10924
Every day avg. increase in deaths: 2402

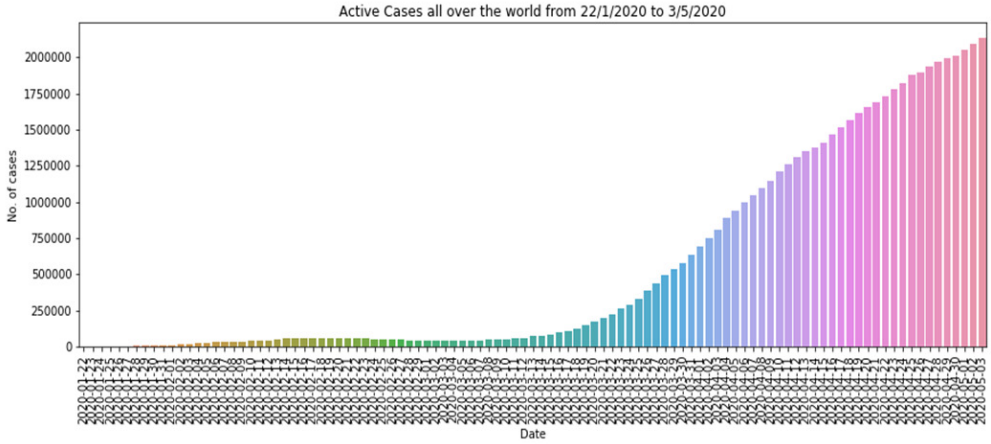

Fig. 1. Active cases all over the world from Feb.' 20.

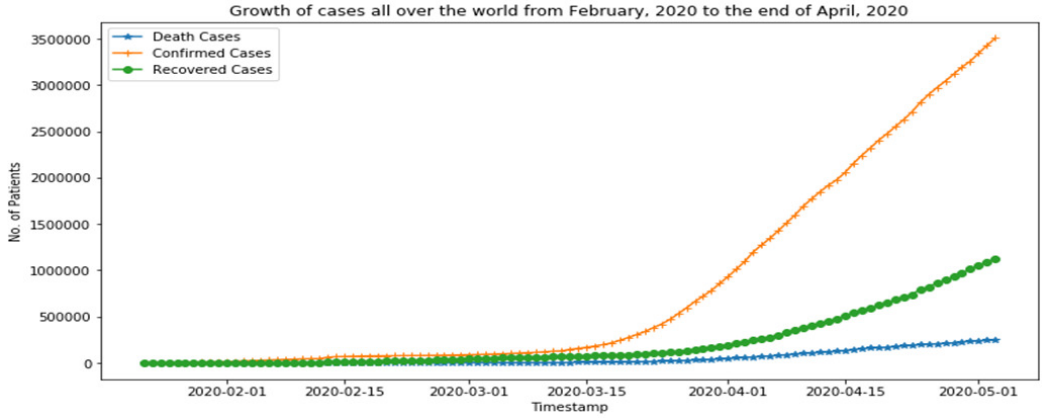

Fig. 2. Growth of cases all over the world from Feb.' 20.

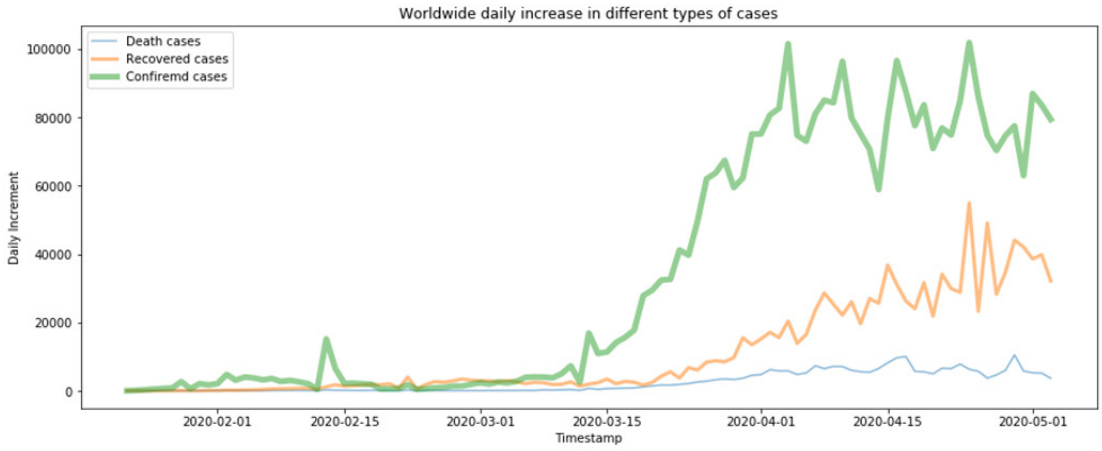

Fig. 3. Daily rise in different type of cases all over the world from February, 2020 onwards.

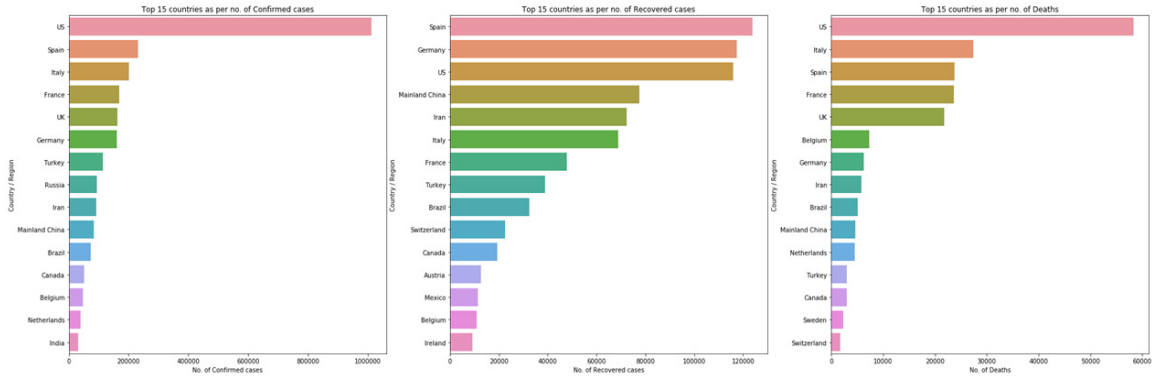

Fig. 4. Statistics of different type of cases of top 15 countries all over the world.

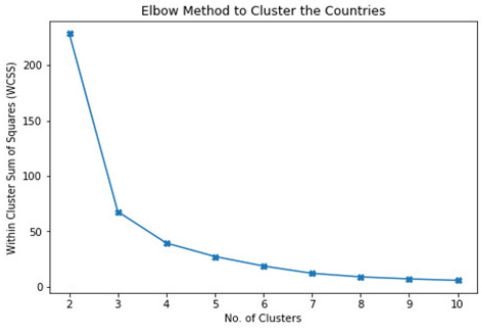

Fig. 5. Elbow method to cluster the Countries.

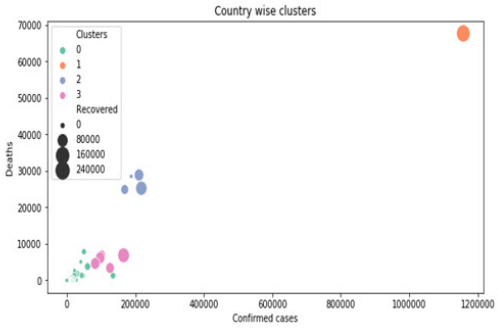

Fig. 6. Country wise clusters.

Table 1: Cluster wise Country names and No. of Countries.

Country/Region	Cluster	No. of Countries
US	1	1
Spain, Italy, UK, France	2	4
Germany, Turkey, Brazil, Iran, Mainland China	3	5
Russia, Canada, Belgium, Peru, India, Netherlands, Switzerland etc	0	179

Country wise rise in confirmed, recovered and deaths is depicted in Fig. 4 for the top 15 countries. It is clear from the figure that the values are different for different countries and United States stands top in the confirmed and deaths, where as in case of recovery of patients it is in third position.

The countries can be grouped based on – Confirmed cases, Recovered cases and Deaths as metrics. By using the "KMeans clustering technique" the countries can be clustered and is shown in the Fig. 5. This shows the "elbow method" to identify number of clusters and it is four clusters. By applying "KMeans clustering" with number of clusters as 4, we got the following country, corresponding clusters and no. of countries. The Fig. 6 presents the "bubble plot" for Confirmed cases, Deaths, Recovered cases on Cluster wise. Table 1 shows the statistics about the clusters and names of the countries and no. of clusters.

IV. IMPLEMENTATION

The COVID-19 blowout has conveyed the world under the edge of damage of human lives. This has made it a greatest prominence to study the transmission progress as early as possible and predict the future options of the transmission. With this objective, we applied Linear Regression, Support Vector Machines (SVM) and Time series forecasting models on the COVID-19 dataset.

These Machine learning methodologies are executed using the Python, to forecast the confirmed, recovered, and death cases worldwide. The forecast will permit us to undertake the needed decisions established on transmission progress such as extending the lockdown period, implementing the hygiene system, providing the daily commodities, etc.

The dataset considered for the analysis is from 22nd January, 2020 to 3rd May, 2020. This dataset is divided into train and test data as 95% of the data from the said date range as train data and remaining 5% is taken as test data. By applying the Linear Regression method and Support Vector Machine (SVM) to the data, we get a model with Root Mean Square Error (RMSE) as,

RMSE for Linear Regression: 1230562.2282881674

RMSE for Support Vector Machine: 371881.172851964

From the Fig. 7, it is clear that the trend of confirmed cases is not linear and Linear Regression is not in the way of actual confirmed cases. At the same time the Fig. 8 indicates, the SVM regression results are far better than the Linear regression.

Time series forecasting methods

Considering the same dataset for time series predicting using Holt's Linear model and Holt's Winder model gives the following RMSE values and the charts are presented in Fig. 9 and Fig. 10.

RMSE for Holt's Linear Model: 11721.579454409777

RMSE for Holt's Winter Model: 14919.507034216318

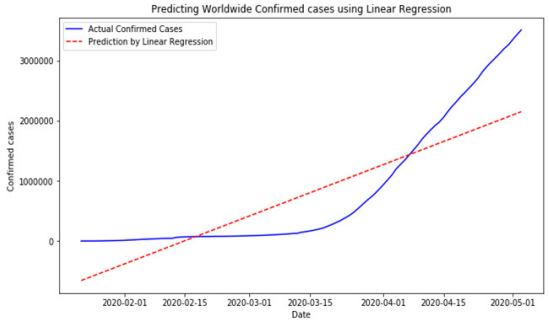

Fig. 7. Prediction of Confirmed cases using Linear Regression.

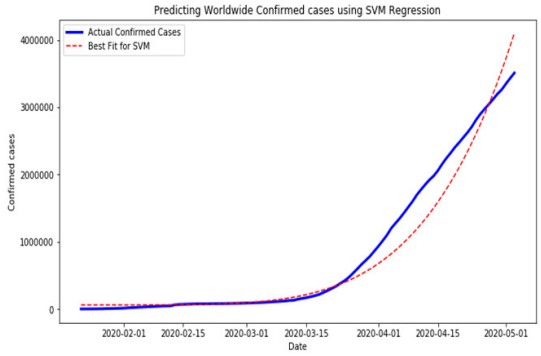

Fig. 8. Prediction of Confirmed cases using Support Vector Machine Regression.

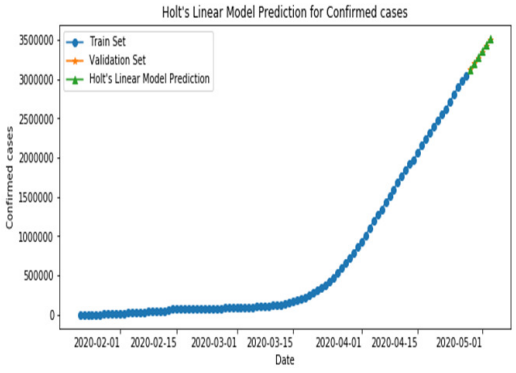

Fig. 9. Holt's Linear Model.

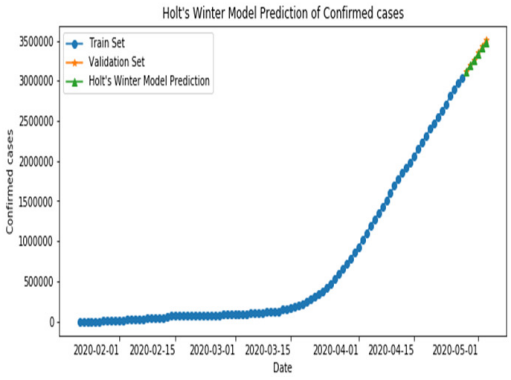

Fig. 10. Holt's Winder Model cases.

ARIMA stands for Auto Regressive Integrated Moveing Average method is implemented in three ways as AR model, MA model and ARIMA model. The RMSE for these are

RMSE for AR Model: 21189.170688733222
RMSE for MA Model: 62374.24547340992
RMSE for ARIMA Model: 53987.87105302296
The Fig. 11 to 13 shows the predictions of the above three models in comparision with acual values.

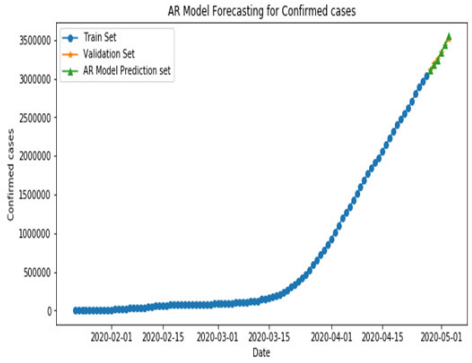

Fig. 11. AR Model forecasting for Confirmed cases.

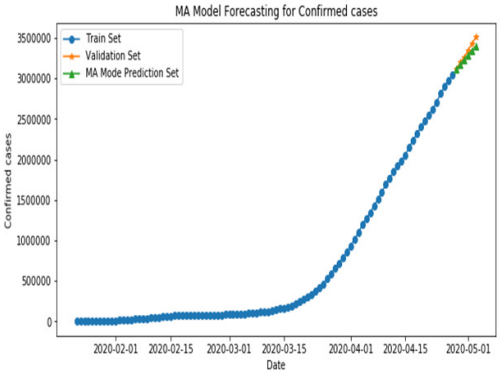

Fig. 12. MA Model forecasting for Confirmed cases.

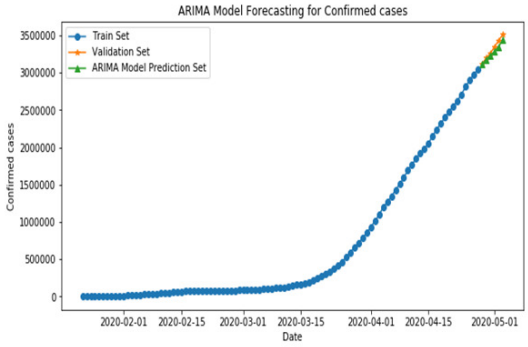

Fig. 13. ARIMA Model forecasting for Confirmed cases.

V. RESULTS ANALYSIS

The Table 2 shows the RMSE values for the seven models – Linear Regression, SVM, Holt's Linear, Holt's Winter, Auto Regressive, Moving Average and ARIMA models implemented in this paper on the COVID-19 data from 22nd February, 2020 to 3rd May, 2020. From the table it is clear that the Holt's Linear model is with lower RMSE value and the predictions are close to the actuals.
The predicted Confirmed cases all over the world for the next five days are shown in the Table 3.
From this it is clear that the change between the real values and values projected by the Holt's model are very less comparing with the other six models.

Table 2: Root Mean Square Error (RMSE) values for the models.

Model Name	Root Mean Squared Error (RMSE)
Holt's Linear	11721.579454409777
Holt's Winter Model	14919.507034216318
Auto Regressive Model (AR)	21189.170688733222
ARIMA Model	53987.87105302296
Moving Average Model (MA)	62374.24547340992
Support Vector Machine	371881.17285196413
Linear Regression	1230562.2282881674

Table 3: Predictions on Confirmed cases around the World for five days.

Dates	Linear Regression	SVM	Holt's Linear Model	Holt's Winter Model	AR Model	MA Model	ARIMA Model
2020-05-04	2176216	4299114	3592733	3553164	3630694	3438569	3481978
2020-05-05	2203704	4508899	3671306	3624640	3746260	3478878	3568568
2020-05-06	2231193	4726910	3749880	3698865	3844089	3514505	3622321
2020-05-07	2258682	4953386	3828453	3770286	3956600	3545298	3655498
2020-05-08	2286171	5188571	3907026	3844566	4092028	3571122	3681622

VI. CONCLUSION

The World is struggling with the COVID-19 virus. General public are lasting their life's in large numbers. Health care and the budget are under heavy pressure. Governments are struggling hard to guard their citizens. But the COVID-19 has showed us a diverse message, a message of interdependence. We were influenced by each other for our health care tools, transportation facilities, amenities, and lastly the vaccines, as and when they are developed.

From the above study a positive observation is that COVID-19 doesn't have very high death rate and recovery rate clearly indicates that this virus is curable. The worrying factor is the exponential growth rate of infected cases. Especially European countries like USA, Italy, UK and Spain are facing problems regarding growth in the active cases and deaths. A positive sign observed from the analysis is that there is some slowdown in the growth of Confirmed and Death Cases in few countries. If any new country appears as new epicenter, the growth of Confirmed cases will increase again.

VII. FUTURE SCOPE

The analysis in this paper is limited to few time series models. This can be extended to other models like Auto ARIMA, Facebook Prophet to improve the accuracy and reduce RMSE.

Conflict of Interest. The authors confirm that there are no known conflicts of interest associated with this publication of this paper.

REFERENCES

[1]. Zhu, N., Zhang, D., & Wang, W. (2020). China Novel Coronavirus Investigating and Research Team. A novel coronavirus from patients with pneumonia in China, 2019 [published January 24, 2020]. *New England Journal of Medicine*.

[2]. Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., & Xing, X. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine, 383*(13), 1199-1207,

[3]. Cohen, J., Normile, D. (2020). New SARS-like virus in China triggers alarm, [J]. Science, Jan 17, 2020, *367*, 234-235.

[4]. Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., & Mulders, D. G. (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro surveillance*, *25*(3), 1-8.

[5]. Chan, J. F. W., Yuan, S., Kok, K. H., To, K. K. W., Chu, H., Yang, J., & Tsoi, H. W. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*, *395*(10223), 514-523.

[6]. Wang, C., Horby, P. W., & Hayden F.G. (2020). A novel coronavirus outbreak of global health concern, [J]. Lancet, Feb 15, 2020, *395:* 470-473,

[7]. Holshue M.L, DeBolt C., & Lindquist S., (2020). First Case of 2019 Novel Coronavirus in the United States, *The New England Journal of Medicine*, Jan 31, 2020, 929-936. doi: 10.1056/NEJMoa2001191.

[8]. Chen, N., Zhou, M., & Dong X. (2020), Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study, [J]. Lancet. Jan 30, 2020, *395*(10223): 507-513.

[9]. Shi, C., Wu, C., Han, X., Xie, Y. and Li, Z., (2016), Machine Learning under Big Data, *6th International Conference on Electronic, Mechanical, Information and Management*.

[10]. Sucharitha, B. Lakshmi, Ch. V. Raghavendran, & B. Venkataramana (2019). Predicting the Cost of Pre-owned Cars Using Classification Techniques in Machine Learning, *International Conference on Advances in Computational Intelligence and Informatics*, 253-263. Springer.

[11]. G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch. Srinivasulu (2019). House Price Prediction Using Machine Learning. *International Journal of Innovative Technology and Exploring Engineering, 8*(9): 717-722.

[12]. Holt, C. C., United States, & Carnegie Institute of Technology (1957). Forecasting seasonals and trends by exponentially weighted moving averages. Pittsburgh, Pa: Carnegie Institute of Technology, Graduate school of Industrial Administration.

[13]. Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science, 6*(3): 324–342.

[14]. Github repository. 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE.

[15]. Domingos S. de O. Santos Júniorab João F.L.de Oliveirab Paulo S.G. de Mattos Neto, (2019). An intelligent hybridization of ARIMA with machine learning models for time series forecasting, Knowledge-Based Systems, *175*(1): 72-86.

How to cite this article: Raghavendran, C. V., Satish, G. N., Krishna, V. and Basha, S. M. (2020). Predicting Rise and Spread of COVID-19 Epidemic using Time Series Forecasting Models in Machine Learning. *International Journal on Emerging Technologies*, *11*(4): 56–61.