IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

COVID-19 Future Forecasting using Machine Learning Model

R. Arunachalam[1], Naveena N[2], Suganthi D[3]
Assistant Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3]
Anjalai Ammal Mahalingam Engineering College, Thiruvarur

Abstract: *Machine learning (ML) based forecasting mechanisms have proved their significance to anticipate in perioperative outcomes to improve the decision making on the future course of actions. The ML models have long been used in many application domains which needed the identification and prioritization of adverse factors for a threat. Several prediction methods are being popularly used to handle forecasting problems. This study demonstrates the capability of ML models to forecast the number of upcoming patients affected by COVID-19 which is presently considered as a potential threat to mankind. Three types of predictions are made by Linear Regression Model such as the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The results produced by the study proves it a promising mechanism to use these methods for the current scenario of the COVID-19 pandemic.*

Keywords: COVID-19, Supervised Machine Learning Model, Linear Regression, Root Mean Square Error, etc.

I. INTRODUCTION

Machine learning (ML) has proved itself as a prominent field of study over the last decade by solving many very complex and sophisticated real-world problems. The application areas included almost all the real-world domains such as healthcare, autonomous vehicle (AV), business, natural language processing robots, gaming, climate modelling, voice, and image processing. ML algorithms' learning is typically based on trial-and-error method quite opposite of conventional algorithms which follows the programming instructions based on decision statements like if-else [1]. One of the most significant areas of ML is forecasting [2], numerous standard ML algorithms have been used in this area to guide the future course of actions needed in many application areas including weather forecasting, disease forecasting, stock market forecasting as well as disease prognosis.

Various regression and neural network models have wide applicability in predicting the conditions of patients in the future with a specific disease [3]. There are lots of studies performed for the prediction of different diseases using machine learning techniques such as coronary artery disease [4], cardiovascular disease prediction [5], and breast cancer prediction [6]. In particular, the study [7] is focused on live forecasting of COVID-19 confirmed cases and study [8] is also focused on the forecast of COVID-19 outbreak and early response. These prediction systems can be very helpful in decision making to handle the present scenario to guide early interventions to manage these diseases very effectively.

This study aims to provide an early forecast model for the spread of novel coronavirus, also known as SARS-CoV-2, officially named as COVID-19 by the World Health Organization (WHO) [9]. COVID-19 is presently a very serious threat to human life all over the world. At the end of 2019, the virus was first identified in a city of China called Wuhan, when a large number of people developed symptoms like pneumonia [10]. It has a diverse effect on the human body, including severe acute respiratory syndrome and multi-organ failure which can ultimately lead to death in a very short duration [11]. Hundreds of thousands of people are affected by this pandemic throughout the world with thousands of deaths every coming day. Thousands of new people are reported to be positive every day from countries across the world. The virus spreads primarily through close person to person physical contacts, by respiratory droplets, or by touching the contaminated surfaces. The most challenging aspect of its spread is that a person can possess the virus for many days without showing symptoms. The causes of its spread and considering its danger, almost all the countries have declared either partial or strict lockdowns throughout the affected regions and cities.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

Medical researchers throughout the globe are currently involved to discover an appropriate vaccine and medications for the disease. Since there is no approved medication till now for killing the virus so the governments of all countries are focusing on the precautions which can stop the spread. Out of all precautions, "be informed" about all the aspects of COVID-19 is considered extremely important. To contribute to this aspect of information, numerous researchers are studying the different dimensions of the pandemic and produce the results to help humanity.

To contribute to the current human crisis our attempt in this study is to develop a forecasting system for COVID-19. The forecasting is done for the three important variables of the disease for the coming 10 days: 1) the number 0f New confirmed cases. 2) the number of death cases 3) the number of recoveries. This problem of forecasting has been considered as a regression problem in this study, so the study is based on some state-of-art supervised ML regression model as linear regression (LR). The dataset has been pre-processed and divided into two subsets: training set (85% records) and testing set (15% records). The performance evaluation has been done in terms of important measure including Root mean square error (RMSE).

II. OVERVIEW OF COVID-19

Coronavirus, the pandemic that is spreading around the world, has uncovered the weakness of human culture to serious irresistible illnesses and the trouble of taking care of this issue in a universally interconnected complex framework. Coronavirus influenced in excess of 100 nations in a range of weeks. As an outcome, the entire human race ought team up to conquer the pestilence as well as sensibly organize to re-visitation of work and creation as per the genuine circumstance of every district and do topographical danger evaluation. Numerous endeavours have been directed to locate an appropriate and quick approach to distinguish tainted patients in a beginning phase.

Subsequent to making chest CT sweeps of 21 patients tainted with COVID19 in China, Guan et al found that CT filter examination included respective pneumonic parenchymal ground-glass and consolidative aspiratory opacities, in some cases with an adjusted morphology and a fringe lung dispersion. Thusly, COVID-19 analysis can be spoken to as a picture division issue to remove the principal highlights of the disease. The sickness brought about by the novel Covid, or Coronavirus Disease 2019 (COVID-19) is rapidly spreading internationally. It has contaminated in excess of 1,436,000 individuals in excess of 200 nations and domains as of April 9, 2020.

III. SUPERVISED MACHINE LEARNING MODEL

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y). In the case that an undefined input instance is given, a supervised learning model can predict. Thus, data sets of input instances and its respective input instances for technique learning by the learning algorithms.

The regressor is used for the regression model. A forecast for unpredictable entrants or test data is then generated in the qualified model. For the development of predictive models, Regression techniques and classification algorithms study method used here.

IV. 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 learning 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. Two variables (x, y) are necessary for the linear regression search. This equation indicates how y is associated with ×, which is called regression.

$$y = \beta 0 + \beta 1x + \varepsilon \qquad (1)$$

$$E(y) = \beta 0 + \beta 1x \qquad (2)$$

This is the linear term for error regression. This error term takes into account the variability between × and y, β0 is the y-intercept and ß1 is the pitch.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

A class mark is specified in the input data set for the purpose of the model × training of the linear regression in the context of machine study. The aim is to find the optimum values for β0 (intercept) and β1 (coefficient) to get the best regression line. The difference between the actual values and the values predicted should be minimum to make sure that this minimising problem is presented:

Minimize1n∑i=1n(predi−yi)2 (3)

Here, g, which is the mean root square of the expected value for y (predi) and y (yi), n is the cumulative number of data points. g is called the cost function.

V. ROOT MEAN SQUARE ERROR

Root mean square error can be defined as the standard deviation of the prediction errors. Prediction errors also known as residuals is the distance from the best fit line and actual data points. RMSE is thus a measure of how concentrated the actual data points are around the best fit line. It is the error rate given by the square root of MSE given as follows

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (4)$$

VI. METHODOLOGY

The study is about novel coronavirus also known as COVID-19 predictions. The COVID-19 has proved a present potential threat to human life. It causes tens of thousands of deaths and the death rate is increasing day by day throughout the globe. To contribute to this pandemic situation control, this study attempts to perform future forecasting on the death rate, the number of daily confirmed infected cases and the number of recovery cases in the upcoming 10 days. The forecasting has been done by using four ML approaches that are appropriate to this context. The dataset used in the study contains daily time series summary tables, including the number of confirmed cases, deaths, and recoveries in the past number of days from which the pandemic started. Initially, the dataset has been pre-processed for this study to find the global statistics of the daily number of deaths, confirmed cases, and recoveries.

VII. STEPS INVOLVED IN PROPOSED METHOD

1. Collect datasets with different parameters from Ministry of India and world meters.
2. Train 80%, validate 10%, and test on 10% of the collected sample datasets.
3. Apply the proposed hybrid model to forecast the Covid-19 data trends.

VIII. CLASSIFICATION OF DATA

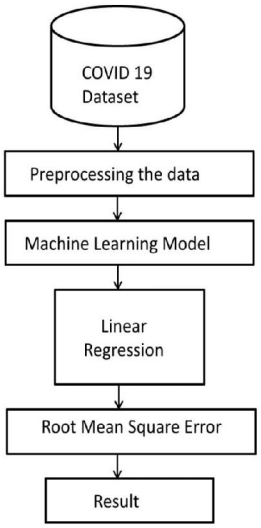

Figure 1: System Architecture

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

The data is classified into three categories:

1. Confirmed cases
2. Recovered cases
3. Deaths

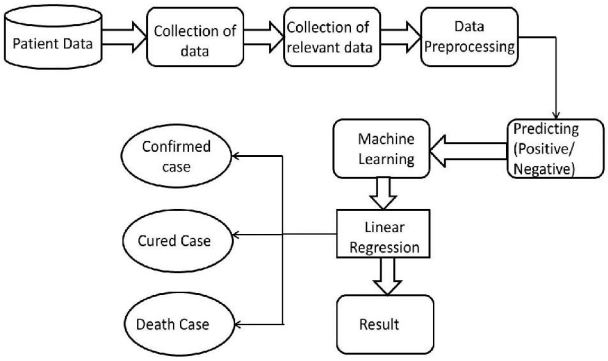

Figure 2: Overall System Flow Diagram

IX. WORKING

Data Preprocessing

The purpose of pre-processing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model.

Data Collection

It's time for a data analyst to pick up the baton and lead the way to machine learning implementation. The job of a data analyst is to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analyzing results with the help of statistical techniques. The type of data depends on what you want to predict. There is no exact answer to the question "How much data is needed?" because each machine learning problem is unique. In turn, the number of attributes data scientists will use when building a predictive model depends on the attributes' predictive value.

Data Formatting

The importance of data formatting grows when data is acquired from various sources by different people. The first task for a data scientist is to standardize record formats. A specialist checks whether variables representing each attribute are recorded in the same way. Titles of products and services, prices, date formats, and addresses are examples of variables. The principle of data consistency also applies to attributes represented by numeric ranges.

Data Splitting

A dataset used for machine learning should be partitioned into three subsets training, test, and validation sets.

Training Set

A data scientist uses a training set to train a model and define its optimal parameters it has to learn from data.

Test Set

A test set is needed for an evaluation of the trained model and its capability for generalization. The latter means a model's ability to identify patterns in new unseen data after having been trained over a training data. It's crucial to use different subsets for training and testing to avoid model over fitting, which is the incapacity for generalization we mentioned above.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, June 2022
Impact Factor: 6.252

X. CONCLUSION

The precariousness of the COVID-19 pandemic can ignite massive global crisis. Some researchers and government agencies throughout the world have apprehensions that the pandemic can affect a large proportion of the world population. In this study, an ML-based prediction system has been proposed for predicting the risk of COVID- 19 outbreak globally. The system analyses dataset containing the day-wise actual past data and makes predictions for upcoming days using machine learning algorithm. The study forecasts thus can also be of great help for the authorities to take timely actions and make decisions It was very difficult to put an accurate hyperplane between the given values of the dataset. Overall, we conclude that model predictions according to the current scenario are correct which may be helpful to under- stand the upcoming situation contain the COVID-19 crisis. This study will be enhanced continuously in the future course, next we plan to explore the prediction methodology using the updated dataset and use the most accurate and appropriate ML methods for forecasting. Real-time live forecasting will be one of the primaries focuses in our future work

XI. RESULT AND DISCUSSION

This study attempts to develop a system for the future forecasting of the number of cases affected by COVID-19 using machine learning methods. The dataset used for the study contains information about the daily reports of the number of newly infected cases, the number of recoveries, and the number of deaths due to COVID-19 worldwide. As the death rate and confirmed cases are increasing day by day which is an alarming situation for the world.

The number of people who can be affected by the COVID-19 pandemic in different countries of the world is not well known. This study is an attempt to forecast the number of people that can be affected in terms of new infected cases and deaths including the number of expected recoveries for the upcoming 10 days. Machine learning model Linear Regression has been used to predict the number of newly infected cases, the number of deaths, and the number of recoveries.

REFERENCES

[1] S. Makridakis, E. Spiliotis and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward", PLoS ONE, vol. 13, Mar. 2018.
[2] G. Bontempi, S. B. Taieb and Y.-A. Le Borgne, "Machine learning strategies for time series forecasting", Proc. Eur. Bus. Intell. Summer School, pp. 62-77, 2012.
[3] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus COVID-19", PLoS ONE, vol. 15, no. 3, Mar. 2020.
[4] G. Grasselli, A. Pesenti and M. Cecconi, "Critical care utilization for the COVID-19 outbreak in lombardy italy: Early experience and forecast during an emergency response", JAMA, vol. 323, no. 16, pp. 1545, Apr. 2020.
[5] C. P. E. R. E. Novel, "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (Covid-19) in China", Zhonghua Liu Xing Bing Xue Za Zhi= Zhonghua Liuxingbingxue Zazhi, vol. 41, no. 2, pp. 145, 2020.
[6] N. C. Mediaite, Harvard Professor Sounds Alarm on 'Likely' Coronavirus Pandemic: 40% to 70% of World Could be Infected This Year, Feb. 2020.
[7] R. Kaundal, A. S. Kapoor and G. P. Raghava, "Machine learning techniques in disease forecasting: A case study on rice blast prediction", BMC Bioinf., vol. 7, no. 1, pp. 485, 2006.
[8] Alsaeedy, A. A. R., & Chong, E. (2020). Detecting Regions at Risk for Spreading COVID-19 Using Existing Cellular Wireless Network Functionalities. IEEE Open Journal of Engineering in Medicine and Biology.
[9] Sear, R. F., Velasquez, N., Leahy, R., Restrepo, N. J., El Oud, S., Gabriel, N., Johnson, N. F. (2020). Quantifying COVID-19 content in the online health opinion war using machine learning. IEEE Access.
[10] Zhang, Y., Li, Y., Yang, B., Zheng, X., & Chen, M. (2020). Risk Assessment of COVID-19 Based on Multisource Data from a Geographical View. IEEE Access.
[11] C. P. E. R. E. Novel et al., "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19) in China.
[12] J.-H. Han and S.-Y. Chi, "Consideration of manufacturing data to apply machine learning methods for predictive manufacturing," in 2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN). IEEE, 2016, pp. 109-113.