JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Proposed System for Prediction of Covid-19 Cases Using Machine Learning

[1]Colin Gonsalves, [2]Dr. Prashant Nitnaware
[1]Student, [2]Project Guide
[1]Department of Information Technology,
[1]PCE, Mumbai, India

Abstract : Covid 19 being the global pandemic, it has become an absolute necessity to make sure the epidemic is controlled and the state of the world with respect to health, finance and livelihood come to a better position. The motive of this research aims to contribute towards achieving this state. Every Local Body/Municipality, of a country, needs to be well equipped with the best technology available. Machine learning will play an important role to help Local Body/Municipality predict the trend of Covid19 cases based on the figures collected in their area. This tool will help the Local Body prepare for medical facilities, Administrative decisions and Citizen lifestyle. Making use of five ML algorithms, namely, Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), FB Prophet, Linear Regression (with FB Prophet) and Random Forest (with FB Prophet). Visualizing the trend on graph, testing the prediction against actual, ruling out the better suited algorithm and using the best of five to predict the coming days.

I. INTRODUCTION

Belonging to a family of coronavirus, Covid 19 came into known in the year 2019. Thus, obtaining the abbreviation Covid 19 for CoronaVirisDeseases 2019. Similar to the rest of the viruses in the family, Covid 19 attacks the respiratory system of the human body. Originating from the Wuhan province in China, it gained momentum and spread across the world. Considering the behaviour and nature of the virus and the rate of spread, the World Health Organization (WHO) declared it as a global pandemic. Along with this pandemic announcement came a set of rules which restricted travel, imposed social distancing and wearing a mask. These steps were imposed to limit or control the spread of virus

India registered its first case on 30th January, 2020 in the state of Kerala. India underwent lockdown to harness the spread of the virus. Registering upto 90 Thousand cases highest in the month of September 2020 and 4 Lakh cases in May 2021 per day, marking the highest peak for the two waves. With availability of vaccines, India initiated the drive in January 2021 which included British Oxford–AstraZeneca vaccine (Covishield), The Indian BBV152 (Covaxin) and Russian Sputnik V vaccine (Emergency use).

Learning about Covid 19, a huge magnitude of data is generated. This data is broadly categorized into state and county data, growth rate variation, covid-19 projections, lab data, disparities and at-risk populations, economic data, mobility, lab data and mortality. Taking a part of the mentioned data available, we can use machine learning techniques to implement certain algorithms to perform predictions under test and actual environment.

Machine learning plays an important role in the medical sector in helping apply predictions based on certain algorithms. These algorithms used depend on how and what the data is. In this report we have tried to understand five machine learning algorithms and how they have been helpful in the past. These algorithms will perform predictions based on the data provided as reference. Algorithms such as Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), FB Prophet, Linear Regression and Random Forest Regressor have proved to be more popular and better with predictions. Random Forest has been more popular with the past research on other diseases such as pneumonia, diabetic, etc. Whereas FB prophet has gained popularity more specific to Covid-19.

As we have seen in the past, the first wave and second wave in India has tested the best effectiveness of our medical facility. Adding additional support to this facility will only provide benefits. Allowing this accessibility, a prediction tool made handy to the local bodies of the state, will help us prepare for medical requirements, administrative actions and citizen awareness.

There would be certain challenges that we expect the Municipality to face but the benefit outweighs the challenges. In the current scenario, the municipality has to depend on the figures available on the internet. The online prediction portal available is performing predictions on the country or state level data. calculation based on prorata is presented on the municipality level. Though this is a workaround, the same cannot be considered as appropriate. Also considering these tools are not easily available in the market and the ones available are expensive. Making such tools available will not only provide support to the municipality but also to the nation indirectly.

Municipality bodies in India do not have easy access to the prediction tools to predict the rise or fall of Covid 19 new cases on a day to day basis. These bodies have to be dependent on figures available on the internet. The drawback of these figures are that they are available at country/state/city level. Very few municipalities have their own tool to perform predictions. Considering these

scenarios, having such tools made available at the local body level in itself will provide a greater impact. Close to correct prediction will help the medical facility, administrative department and citizens be prepared for the coming day. Medical facilities such as availability of hospital beds, oxygen cylinders, medical personnel, etc. Administrative department can focus on implementing lockdows accordingly. Individual citizens can be made aware to take necessary actions. Taking care at the local level will add to the hierarchy of City, State, Country and World. Hence making such a facility available is indirectly protecting the human kind.

Considering the algorithm that will be used to perform the prediction:

Random Forest Regressor has shown better results in the past, comparatively. Random forest is a learning method that is supervised. It creates a "forest" out of an ensemble of decision trees, which are generally trained using the "bagging" approach. The bagging method's basic premise is that combining several learning models improves the final outcome.

ARIMA (autoregressive integrated moving average) is a statistical analysis technique that use time series data to better understand a data collection or forecast future trends. If a statistical model predicts future values based on previous values, it is called autoregressive.

The control flow of an LSTM is comparable to that of a recurrent neural network. It processes data and passes information on as it moves along. The processes within the cells of the LSTM vary. The LSTM uses these processes to remember or forget information.

FBProphet aims to fit several linear and nonlinear time functions as components using time as a regressor. By default, FBProphet uses a linear model to fit the data, but it may be switched to a nonlinear model (logistics growth) using its parameters. There are a sufficient number of missing or outlier data points.

Linear regression is a supervised learning-based machine learning technique. Linear regression is a mathematical technique for predicting the value of a dependent variable (y) based on the value of an independent variable (x). As a result of this regression approach, a linear connection between x (input) and y (output) is discovered (output).

II. LITERATURE SURVEY

For the purpose of this research, past research was referred to as providing a base and compounding effect to add better value. The search was categorized to look for research work done on Machine Learning, Prediction algorithms and Covid 19. Machine learning research in to health care include Pneumonia Detection by Chouhan V. (2020), Diseases Classification by Sharmila S. (2017), Chest diseases diagnosis by Er O. (2010), Role of machine learning to predict the outbreak of covid-19 by Tiwari U.K. (2020) etc. Diseases prediction research include Building predictive models for MERS-CoV infections by Turaiki I. (2016), Disease prediction using machine learning by Shirsath S.S. (2018), prediction of the epidemics trend of COVID-19 by Yang Z.F. (2020), Predicting active pulmonary tuberculosis by El-Solh A.A. (1999), Predicting the Growth and Trend of COVID-19 by Tuli S. (2020).

When we consider past works that have been performed on the algorithms, a summary for the same is provided below:

ARIMA:

This is to stress the importance of predicting the number of confirmed instances rather than the time of the forecast. Minimum, maximum, and high variability were divided into five categories. The categories were split into 19 instances based on empirical data analysis. The auto regressive integrated moving average (ARIMA) model is used to estimate the total number of COVID-19 confirmed cases, which is then compared to the actual number of confirmed cases. Forecasts may correctly detect diminishing and growing trends using group and case-by-case prediction. To stop COVID-19 from spreading further, the government must act quickly and firmly. This research will aid the government in responding to a potential spike in confirmed cases in a methodical manner.

LSTM:

In theory, an LSTM can recall long-range information and track the different properties of the text it is currently processing using its memory cells. For example, writing gadget cell weights that allow the cell to keep track of whether it is inside a quoted string is a straightforward exercise. A cell, an input gate, an output gate, and a forget gate make up a standard LSTM unit. Over self-assertive, the cell recollects values. The evolution of data relating to the cell is managed by periods and the three gates. The vanilla variety will be alluded to throughout the rest of this chapter, as it is the most well-known LSTM engineering.

FB Prophet:

Using the FB prophet model, the researchers were able to forecast the active, death, and healed rates of COVID 19 in Algeria for a future period of 35 days. The active rate and death rate both decreased during the next several days, but the cured rate increased. The rates of active, cured, and death are predicted to be 19.7%, 78.85%, and 2.55 percent, respectively. These findings show the significance of the FB prophet model in COVID-19 prediction, which may aid national authorities in implementing the most effective preventative measures.

Linear Regression:

The score of the model R2 tends to be 0.99 and 1.0 in this comparison of Linear Regression and Multiple Linear Regression models, indicating a good prediction model to anticipate the next coming days active cases. If the current trend continues, the projected value of 52,290 active cases in India and 9,358 active cases in Odisha for the month of August is anticipated using the Multiple Linear Regression model as of July.

Random Forest Regressor:

On the COVID-19 patient dataset, the paper provided a model that implemented the Random Forest technique and was boosted by the AdaBoost algorithm, with an F1 Score of 0.86. The Boosted Random Forest method was identified to offer good predictions even on unbalanced datasets. According to the statistics evaluated in this study, death rates among Wuhan locals were greater than non-natives. In addition, male patients died at a higher rate than female ones. The majority of those afflicted are between the ages of 20 and 70.

III. EXISTING SYSTEM AND PROBLEM STATEMENT

3.1 Existing System

The current method of prediction is conventional if we refer to the municipalities methods. We interacted with a few of the municipality corporators in VVMC. The method of prediction can be elaborated as a "Comparative" method. Taking as an example: https://covid19.healthdata.org/ provides prediction of the covid 19 daily counts on a country level. These figures are brought down and compared with figures available in their area.

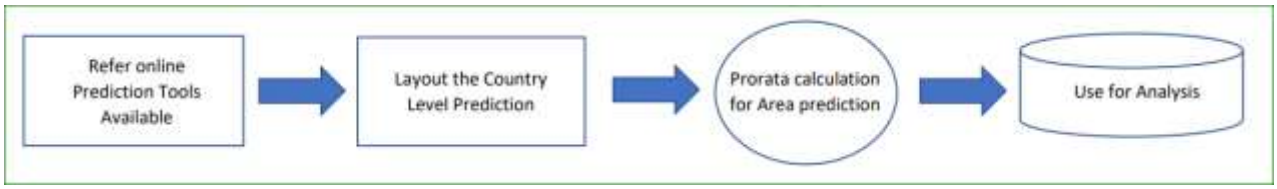

Fig 3.1 Existing System

Elaborating this, let us understand with an example. The data presented for the country is collected from every municipality. The cumulative of city and state data provides the count of the country. These counts are presented in front of the nation for understanding the current state. There are many online portals available that provide data visualization to demonstrate the trend underway. Some of these portals amso provide predictions based on the data available. Above is the mentioned example for the prediction portal. The prediction calculation at municipality level is performed in a reverse chronological order. Hence when the production of the country is performed as "X" the same is calculated against the country data. The percentage increase of decrease is calculated and the same is applied at the municipality data. The same is calculated for the next 15 to 20 days data and presented for analysis. Though this process is a workaround, the same has been used for a while. As the tool itself is not available, the local body has to be dependent on such calculations.

If such is the scenario, the data is not fully reliable because the prediction may not be completely in ratio with the country level prediction. Given this possibility, we do propose this application to perform predictions on the actual Area data which may provide better and dependable predictions then the conventional method.

3.2 Problem Statement

Considering the current system, the method used cannot be termed as appropriate. Though it is a workaround to fill the gaps present. Hence the primary agenda is availability. To make this tool available to the municipality as they do not have easy access to perform prediction on the data available with respect to their own area.

Data visualization itself takes time, as the data needs to be furnished and calculated. This visualization itself has its own challenges when it comes to performing comparison. Currently municipalities do not have a scope to choose which algorithms best suit their area calculations. This can be resolved by providing options to perform operations based on different algorithms and compare the prediction of each. The best suited one can be used to perform the actual coming days prediction.

IV. PROPOSED SYSTEM

The proposed system is based on a machine learning framework. Implementing five machine learning tools to learn, identify and predict the pattern of covid 19 cases being registered. This system intends to implement ARIMA, LSTM, FB Prophet, Linear Regression and Random Forest Regressor.

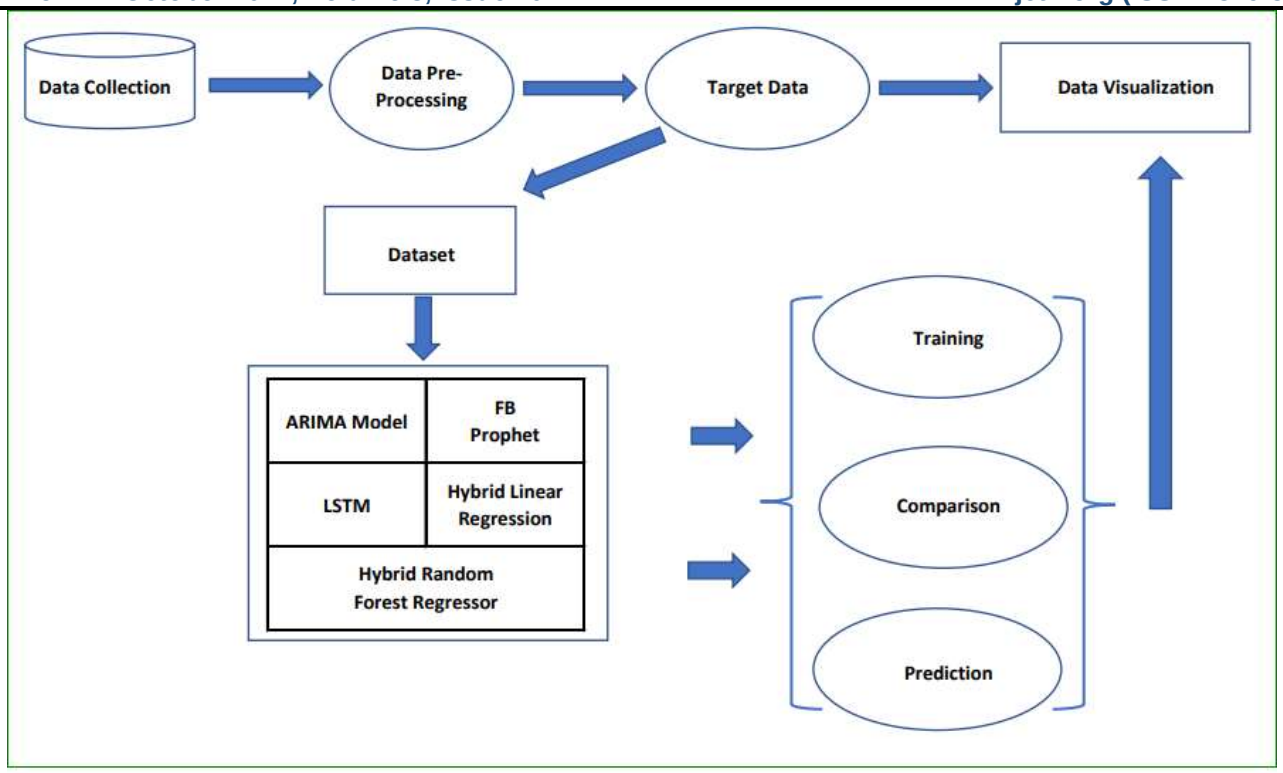

Fig 3.2 Proposed System

4.1. Data Collection

The process for collecting the data for the required operation from one or more sources is termed as data collection. The raw collected data in itself may be in one or multiple formats. All these formats of data are collected and placed ready. These sets of data formats are also known as datasets. For every machine learning system, data is the key ingredient. This will act as the input for the algorithm and relative to it the output will be derived.

We intend to use the data from the below mentioned source:

https://api.covid19india.org/

Covid19India.org makes the data available in a comma delimited file. Depending on the date range, the number of files are downloaded. The data is made available at City/Municipality level. Mumbai (BMC), Thane (TMC) and Palghar (VVMC) are the main raw data points that will be used to extract the

New Cases for the Day

New Deaths for the Day

New Recovered for the Day

In the next stage we will understand how the data will be furnished.

4.2. Data Pre-Processing

For processing the file, below is the accepted format with fields as:

ID,Day,Month,Year,Area,New,Recovered,Deceased

Pre-processing of files involves making the data ready and furnished for operational use. Defining this format is what can be called the main dataset for the prediction. The furnishing of the dataset will involve manual work or an automated script to perform the task. For the current scope, we will proceed with manual intervention to prebate the dataset. Explaining the fields:

ID : Unique Identity for the record

Day: Day of the Month

Month: Month of that Year

Year: Year of the Day and Month

(Bringing together to form dd/mm/yyyy)

Area: Name of the Municipality

New: Total New Patient Count for that date and Area

Recovered: Total Recovered Patient Count for that date and Area

Deceased: Total Deceased Patient Count for that date and Area

4.3. Target Data

Based on the data collected and furnished, it will be used to be imported. Data import necessarily requires the pre-set format. This allows the application to proceed with the functionality and the values that each of the fields hold. This imported data will be confirmed and taken for processing as per further requirements.

Supported format of file will be a comma separated (csv) file which will act as a read only database. The visualization depends on how the data is read by the instructions defined. These instructions we will understand further in detail in our next segment under Training Set and Prediction.

4.4. Data Visualization

4.4.1 Training Set

Data visualization is defined as presenting the data in an elaborative, readable and understandable way. The collected data is available in a comma separated which is readable but not quite explanatory. Hence for demonstration purposes, data visualization tools are used to make them understandable in public or closed forums.

Training Set is a part of data visualization. This can also be termed as a testing phase to verify the comparison and effectiveness between mentioned algorithms. We propose to use five algorithms in the training set helping us understand the effectiveness of each of the five algorithms.

Autoregressive Integrated Moving Average (ARIMA)
Long Short-Term Memory (LSTM)
FB Prophet
Linear Regression Hybrid (using FB Prophet)
Random Forest Regressor Hybrid (using FB Prophet)

The training set module is used by the data analyzer to make sure that the data is well tested and the prediction may work as per the expected. This provides the insights to tweak the algorithm pr classify based on the prediction vs actual. For this application we will work based on an 80:20 ratio. This will allow the application to use 80 percent of the sata as base reference and perform the next 20percent prediction. The 20 percent prediction will be compared with the actual values to verify now the algorithm is performing.

These selections may vary on multiple factors. but more importantly it depends how the actual vs prediction is portrayed. The data analyzer will then use the selection to actually predict the future trend. We shall understand the prediction module better in the next section.

4.4.1 Prediction

Prediction module and training dataset module have a dependency. The prediction module depends on the data analyzer to use the best effective algorithm out of the five in the training set. The selected is the one that will perform the prediction based on the instructions set. Even though the prediction module will only be used for one algorithm, we will provide the facility to verify the other algorithms to be used as reference.

The prediction module will make use of the same algorithms as listed below:

Autoregressive Integrated Moving Average (ARIMA)
Long Short-Term Memory (LSTM)
FB Prophet
Linear Regression Hybrid (using FB Prophet)
Random Forest Regressor Hybrid (using FB Prophet)

V. HARDWARE AND SOFTWARE REQUIREMENTS

Hardware Requirements

Processor	: Intel(R) core(TM) i3-7020U CPU@ 2.30GHz
Installed memory (RAM)	: 8.00 GB
System type	: 64-bit OS, X64-based Processor
Input device	: Standard Keyboard and Mouse, no Pen and Touch are required

Software requirements

Operating System	: Windows
Technology	: Anaconda for Python Environment, Streamlit Framework
Database	: Flat File

VI. CONCLUSION

Gathering knowledge on past research more specific to Machine Learning, Prediction Techniques used in the past and Covid 19 we can build a system that will make use of machine learning prediction techniques to predict the trend of Covid 19 cases that will get registered in the respective local body. Making use of five ML algorithms, namely, Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), FB Prophet, Linear Regression (with FB Prophet) and Random Forest (with FB Prophet) the analyst can perform comparison and use the better algorithm to perform prediction. Adding to boost the local body is indirectly contributing to strengthening the country to help prepare the medical and administrative departments.

REFERENCES

[1] Quinlan R. Morgan Kaufmann Publishers; San Mateo, CA: 2014. C4.5: Programs for Machine Learning.
[2] Turaiki I., Alshahrani M., Almutairi T. Building predictive models for MERS-CoV infections using data mining techniques. J. Infect. Public Health. 2016;09:744–748.
[3] Chouhan V., Singh S.K., Khamparia A., Gupta D., Tiwari P., Moreira C., Damaševičius R., De Albuquerque V.H.C. A Novel transfer learning based approach for pneumonia detection in chest X-ray images. Appl. Sci. 2020;10:559.
[4] Sreeja S., Bhavya L., Swamynath S., Dhanuja R. Chest x-ray pneumonia prediction using machine learning algorithms. Int. J. Res. Appl. Sci. Eng. Technol. 2019;07(04):3227–3230.
[5] Kose U., Guraksin G.E., Deperlioglu O. 2015. Diabetes Determination via Vortex Optimization Algorithm Based Support Vector Machines: Medical Technologies National Conference; pp. 1–4.
[6] Sharmila S., Dharuman C., Venkatesan P. Disease classification using machine learning algorithms – a comparative study. Int. J. Pure Appl. Math. 2017;114(06):1–10.
[7] Shirsath S.S. Disease prediction using machine learning over big data. Int. J. Innov. Res. Sci. 2018;07(06):6752–6757.

[8] Er O., Yumusak N., Temurtas F. Chest diseases diagnosis using artificial neural networks. Expert Syst. Appl. 2010;37(12):7648–7655.
[9] Yang Z.F., Zeng Z.Q., Wang K. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J. Thorac. Dis. 2020;12(3):165–174.
[10] Müller M.A., Corman V.M., Jores J., Meyer B., Younan M., Liljander A., Bosch B.J., Lattwein E., Hilali M., Musa B.E., Bornstein S. MERS coronavirus neutralizing antibodies in camels, Eastern Africa, 1983–1997. Emerg. Infect. Dis. 2014;20(12):2093.
[11] El-Solh A.A., Hsiao C.B., Goodnough S., Serghani J., Grant B.J. Predicting active pulmonary tuberculosis using an artificial neural network. Chest. 1999;116(04):968–973.
[12] Tiwari U.K. Role of machine learning to predict the outbreak of covid-19 in India. J. Xi'an Univ. Archit. Technol. 2020;12(4):2663–2669.
[13] Makridakis S., Wakefield A., Kirkham R. Predicting medical risks and appreciating uncertainty. Foresight Int. J. Appl. Forecast. 2019;52:28–35.
[14] Tuli S., Tuli S., Tuli R., Gill S.S. Internet of Things; 2020. Predicting the Growth and Trend of COVID-19 Pandemic Using Machine Learning and Cloud Computing.
[15] Punn N.S., Sonbhadra S.K., Agarwal S. MedRxiv; 2020. COVID-19 Epidemic Analysis Using Machine Learning and Deep Learning Algorithms.
[16] Jia L., Li K., Jiang Y., Guo X. 2020. Prediction and Analysis of Coronavirus Disease 2019. arXiv Preprint; p. 05447. arXiv:2003.
[17] Kalipe G., Gautham V., Behera R.K. IEEE; 2018. Predicting Malarial Outbreak Using Machine Learning and Deep Learning Approach: A Review and analysis; International Conference on Information Technology (ICIT) pp. 33–38.
[18] Sanjay Sharma discusses about How predictive models can aid in the battle against COVID-19. Available from: https://home.kpmg/in/en/home/insights/2020/04/how-predictive-models-can-aid-in-the-battle-against-covid-19.html. (Accessed 30 April 2020).
[19] Bohdan M.P. vol. 04. 2019. (Machine-Learning Models for Sales Time Series Forecast: MDPI). (15)
[20] Santosh K.C. AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data. J. Med. Syst. 2020;44:93.