Machine Learning Based Predictive Model for Coronavirus Pandemic

Supreet Singh\textsuperscript{1}, Lavanya Sharma\textsuperscript{2*}, Birender Kumar\textsuperscript{3}
\textsuperscript{1,2}Amity University Noida, Uttar Pradesh, India
\textsuperscript{3}ABES Engineering College, Ghaziabad and NIU Noida, Uttar Pradesh, India

singh.supreet14@gmail.com, shm.lavanya@gmail.com, bk.uptu@gmail.com

Abstract: Technological advancements have a rapid effect on each domain of day today life, whether it is medical domain or any other domain. Artificial intelligence has shown promising outcomes in emergency department through its decision making by investigating the data. COVID-19 has affected many countries across the globe in a matter of no time. Citizens all over the world are susceptible to its consequences in future. Covid-19 being a contagious disease needs a program that will make it easier for the tracking, the spread and preventing the future cases. Proposed algorithm that can predict on-going Covid-19 positive cases in India. For experimental analysis machine learning algorithm has been coded to make future predictions.

1. Introduction

This year the world faced an invisible giant, in the face of a virus. Even though this virus is microscopic, but is responsible for the death of over a million of people all around the world. Coronaviruses are the group of viruses that belong to the Nidovirales order, which includes Coronavirinae, Arteriviridae, Mesoniviridae and Roniviridae. Coronavirus spreads through contact and can be asymptomatic in some people, this makes the spread easier and quicker as the patient who is not showing any symptoms will not get checked. When something contagious gets out in the world, the biggest challenge that the country or the Government has to face is prediction of the worst-case scenario, and finding out how many people are affected by the disease. Machine learning models can be used for prediction, they are not necessarily accurate all the time, when we are tracking something at this large scale, but they are very helpful when it comes to painting a picture and estimating resources the government will need to tackle the problem.

It was December of 2019, when the cases of novel Coronavirus first appeared in the Wuhan city of China\cite{1}. The virus was reported to the World Health Organization on 31\textsuperscript{st} December 2019. The virus was named COVID-19, by W.H.O.\cite{1}. This virus comes from the family of SARS. W.H.O. declared this outbreak as a global emergency\cite{22-28} and mentioned that this virus is highly contagious and it has been transmitting via the respiratory tract. This virus does not have the same effect on everyone as there are many people who are asymptomatic and hence these people do not get detected until and unless they go through the PCR tests. On
the other hand this virus shows effect on the person in 2 – 14 days, the people who have pre-existing medical conditions are the worst affected. The symptoms range from dry cough to severe cases of dyspnea[3, 4,25,28,29]. No treatments have been discovered so far and the people suffering from the disease are advised to rest and have recovered all by themselves but for the worst cases the people have been admitted to hospital and their vital signs were observed and they were given artificial oxygen.

Machine learning algorithms are mathematics and statistics driven and are targeted to make sense of the data that is being sent through the formula that has been derived or created to get the output. In this report it will be shown how a model can be made for predicting the cases of coronavirus that can be positive in India, in this report real time data has been used. Simple linear regression algorithm will be used[30,32,33,34,35].

2. Machine learning

Machine learning[1] is a field of mathematics and computer science. The concepts of statistics, calculus, probability and linear algebra are used to process and analysis the data and if required make predictions with the data they are given. The use of machine learning has become more common in recent years and will continue to grow in the future too. This was the commercial applications of the machine learning this field is being used tremendously in the areas of research in the medical field and being used to solve and work on the mathematical equations and experiments[25,27,30]. There are 2 kinds of machine learning: Supervised learning is used to predict the outcome of the input and it does this by using the algorithm that is being used, the algorithm used is coded by the person and the algorithm relies heavily on the mathematical side of machine learning. Second type of learning is Unsupervised learning which is a kind of machine learning, where there is no known output, and nobody instructs the learning algorithm. Supervised learning often speeds up the task it is working on with time. This learning is of two types: Firstly, Classification that is used to to predict the class label where, the data point will lie, in many cases the classification is binary and in some cases the classes are multiclass where the algorithm has to classify between multiple class. Secondly, Regression which is employed when the next value is to be predicted what the next outcome will be with the given sets of inputs. The regression can use different algorithms, such as Linear regression, Logistic Regression and K nearest neighbour regression [30,31].

3. Proposed Model

In this proposed model, python 3.7 is used that comes with a lot of inbuilt libraries and the libraries that have been used extensively as numpy, math, pandas, matplotlib, requests and json. Json and request library was used to extract the data from the api. Numpy and pandas were used to clear the data. Matplotlib was used to plot the graphs. Math is used for simplifying mathematical operations. Data is main part of a supervised algorithm. Since this model is a supervised so the data the real time data will be collected from the API. The link to the API that will be used is, link: 'https://corona.lmao.ninja/v2/historical/India?lastdays=50', (API) to get the data. In order to request data, python’s library known as request and JSON will come into play. Pandas library will be used to program the data and automate the process of data updation, extraction and storing. The JSON library gives the format for viewing the data, that is recieved by the request library in json format. The json format is in form of a dictionary and the paramteres can be selected using the keys of the dictionary. Pandas library is used to
convert all the extracted data in form of a dataframe. Dataframe can store the values of multiple lists in form of columns. Dataframe have rows, columns and indices that make it easier and accurate when it comes to accessing these values in the later program. Pandas stores the dataframe in the form of a file, which can later be imported as the work with the model begins.

```
Algorithm 1: Data extraction from an API

Procedure

url ← 'link_of_the_api'

r ← request.get(url)

Determine the key values that are needed to be used from json response. The keys and the values are taken from that response for later use.

Set p1, p2, p3, p4 ← parameter1, parameter2, parameter3, parameter4

P1 = json(load())[p1]
P2 = P1['p2']
P3 = P1['p3']

Make empty lists

set cases, deaths, everyday_cases, everyday_log ← [], [], [], []

For each case in P2:

Append case to cases

Loop ends

For each death in P3:

Append death to deaths

Loop ends

Set i ← 0 and j ← i+1

While check i is less than sizeof (cases)-2 and j is less than sizeof(cases)-1:

append(cases[i] -cases[i]) to everyday_cases

Increment i and j

While loop ends

For each case in everyday_cases:

Append log(case) to everyday_log

For loop ends

Corona = pandas.DataFrame(‘column1’: cases, ‘column2’: deaths, ‘column3’: everyday_cases, ‘column4’: everyday_log)
```

Figure 1: Data extraction

When the data has been extracted and stored into the dataframe and stored as a CSV file. In order to makes sense of the data that has been extracted, the graphs are various plots are made to get the overall idea of the data being dealt with.
A plot of cases and days is made to view how the cases are growing over days. The data comes out to be exponential and that will not be suitable for applying Linear regression model.

Before the model can be applied the values of the graph as shown in Fig.2 have to be made compatible with the application of Linear regression model. In order to do this logarithms of the values is used, as the values are increasing exponentially and so log function which is the inverse of exponential function is used to make it linear as shown in equation 4.

\[ f(x) = e^x \] (1)

When we apply log() to both the LHS and RHS, the equation as shown in Eq.2

\[ \log(f(x)) = x \log(e) \] (2)

Since log has the base e, so:

\[ \log(e) = 1 \] (3)

Hence the final Eq. 4 after applying log to the function is

\[ \log(f(x)) = x \] (4)

The values after this is done are now linear and Linear regression model can be applied.
Now the cases have been changed to logarithmic form as shown in Fig. 3. and the graph now becomes linear and suitable for application of the model. Since a straight line has to be drawn through the datapoints of our model, so this dataset is suitable.
There is a module in python which is known as Scikit-learn and in programming terms it is known as sklearn [3]. The model that is used is a simple linear regression. Real time data is used in the program to make predictions.

1. Linear regression: - Linear models are the models which are used to predict what will come next after the model has been trained. They are called linear models because they are getting inputs using the linear functions[30-35].
1.1. Linear regression: Linear regression is the model that uses two parameters to predict the next outcome of test data point. Linear regression relies on the 'slope' of the line and the intercept of the 'y' axis as shown in Eq. 5. This linear regression works the best when the training dataset is growing linearly and there are minimum outliers.

\[ Y \approx \beta_0 + \beta_1 \times X \]  

(5)

Let \( Y \) be the prediction made made by the model and the true value of the model happens to be \( 'y' \).

Then \( 'e' \) is the value of the prediction as shown in Eq.6 is the deviation from the true value. This deviation represents the residual value. The validity of a model can be expressed as residual sum of square(RSS models.

\[ RSS = e^2 + (e_1)^2 + ... + (e_n)^2 \]  

(6)

The least square approach depend on \( \beta \) and \( \beta_0 \) to minimise the residual sum of square values. As shown in Eq. 7 and Eq. 8.

\[ \beta = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{\sum (x_i - \mu_x)^2} \]  

(7)

\[ \beta_0 = \mu_y - \beta_1 \times \mu_x \]  

(8)

The data for this model has increased exponentially, since this disease is very contagious, here is the graph showing how the cases are spreading across India, the graph is curved and for this dataset it would have been best to use logistic regression, but instead there is always a way to make this exponential graph linear, is by using the log() function.

4. Experimental Analysis

The model uses the Linear regression, that could not be applied on an exponential form of data. This exponential data will confuse the program and the prediction generated will be wrong. The data had to be converted to a form that would not make the predictions inaccurate. The data was made linear by using the logarithm operation. The new values generated after applying logarithm values were now used to calculate the slope and intercept of the data. The slope and intercept are those two values that will be used to make the predictions. The prediction will be shown on the ‘y’ axis of the graph and, using the slope of the line and intercept and input the value in the ‘x’ variable, the prediction will be made.

Linear regression[4] relies on the classic equation of line to make predictions for the future data points. The equation of line is shown in Eq.(9):

\[ y = \text{slope} \times x + \text{y-intercept} \]  

(9)

This is only for two axes, though, if the data is spread across multiple axes then the slope of line in those axes is included in the equation too. The slope of line in linear regression model is referred to as weights. The predicted value hence becomes the weighted sum of all the values across the axes. Finding the slope and intercept of the dataset. The equation for calculation of slope and intercept is shown in Eq 10 and 11: -

\[ \text{Slope} = ((\mu_x \cdot \mu_y) - \mu(x \cdot y))/(\mu_x \cdot \mu_y) - \mu(x \cdot x) \]  

(10)
\[
\text{Intercept} = \mu_y - \text{slope} \times (\mu_x) \quad (11)
\]

Once the slope and intercept is calculated then make a regression function to make the predictions of the data set. This dataset can be an array of numbers we want to predict.

The regression function[5] uses the slope and the intercept to and gives out a line that is most suited for the dataset. This line goes through the middle of the graph where maximum values lies or the value is the mean of the dataset.

The prediction formula is shown in Eq. (12) and is used to predict the possible cases in future.

\[
y = \text{slope} \times (x) + \text{intercept} \quad (12)
\]

![Figure 5 Regressive model](image)

The value of \( x \) here is the input given, and the \( y \) is in accordance with the slope and intercept calculated in Fig (5) above. Slope calculated for this data is given in Eq. 13 and intercept in Eq 14.

\[
\text{Slope} = 0.06417078172445888 \quad (13)
\]

\[
\text{Intercept} = -0.32836667521493523 \quad (14)
\]

This can be used to make predictions by simply appending to this list. The answer here will come in logarithm so, it is necessary that need to be converted to integers using antilog and hence get the answer in log.
The blue dot in the prediction in Fig. 6 is for the 69th from the Coronavirus outbreak in India, the answer is coming in Logarithm so in order to get the answer in integer, antilog is applied on the value. The blue dot's value is 4.099417263772727 which in integer value comes out to be 12573 people diagnosed per day.

5. **Advantages and Disadvantages of Proposed Model**

Though this algorithm was able to make prediction with the given data, but these predictions will be inaccurate and wrong when the cases begin to drop, as the line will always have a constant slope and travel without bending hence, it will not be able to predict when the cases will start dropping. It is always very difficult to make a predictive model in such cases, as I have not considered a lot of other factors that play a vital role in Coronavirus spread. This is the major limitation of this project.

5.1. **Advantages**

1. This model has minimized the error value and the predicted line is not far from all the points, and the there are no outliers.

2. Since the data was taken from an API, there was no need to worry about about the data having NULL values and the data hence this reduced the work to fill in those NULL values.

3. Since the data was showing exponential growth so linear regression model will show continuous growth and accuracy as long as the values keep on increasing.

5.2. **Disadvantages**

1. The disadvantage of this model is that since, the prediction is being made by using a linear regression model, this model will start to show inaccuracy as the number of cases start to decrease and make the next predictions highly inaccurate.

2. The above scenario will also generate outliers and hence increasing the overall error of the model.
6. Conclusion

This paper presents a proposed algorithm that can predict on-going Covid-19 positive cases in India. In this paper an algorithm is developed that will predict the number of cases, in coming days. Also learning how to clean the data we are working with, and making a model that works. Methodology to be adopted: The proposed model is having three steps: First step involves gathering of data that we will be working with, the second step involves the cleaning of the data that we are using, in the final step, a model is proposed that will give correct predictions. The model was coded on Python 3.7 which a very popular programming language and used various modules of Python. Covid-19 being a contagious disease needs a program that will make it easier for the tracking, the spread and preventing the future cases.

References

1. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020) A new coronavirus associated with human respiratory disease in China. Nature 44(59):265–269
2. Medscape Medical News, The WHO declares public health emergency for novel coronavirus (2020) https://www.medscape.com/viewarticle/924596 [Accessed on 10 Apr 2020]
3. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J, Yu T, Zhang X, Zhang L (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 395(10223):507–513.
4. World health organization: https://www.who.int/new-room/g-a-detail/q-a-coronaviruses#/?text=symptoms. [Accessed on 10 Apr 2020]
5. Wikipedia coronavirus Pandemic data: https://en.m.wikipedia.org/wiki/Template:2019%28%29%20_coronavirus_pandemic_data. [Accessed on 10 Apr 2020]
6. Khanday, A.M.U.D., Amin, A., Manzoor, I., & Bashir, R., “Face Recognition Techniques: A Critical Review” 2018
7. Kumar A, Dabas V, Hooda P (2018) Text classification algorithms for mining unstructured data: a SWOT analysis. Int J Inf. Technology.
8. Verma P, Khanday AMUD, Rabani ST, Mir MH, Jamwal S (2019) Twitter Sentiment Analysis on Indian Government Project using R. Int J Recent Tech Eng.
9. Chakraborti S, Choudhary A, Singh A et al (2018) A machine learning based method to detect epilepsy, Int J Inf Technol 10:257–263.
10. Sarwar A, Ali M, Manhas J et al (2018) Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. Int J Inf Technol.
11. Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M (2020) Mapping the landscape of artificial intelligence applications against COVID-19.
12. Wang L, Wong A (2020) COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 Cases from chest radiography images.
13. Yan L, Zhang H-T, Xiao Y, Wang M, Sun C, Liang J, Li S, Zhang M, Guo Y, Xiao Y, Tang X, Cao H, Tan X, Huang N, Amd A, Luo BJ, Cao Z, Xu H, Yuan Y (2020) Prediction of criticality in patients with severe covid-19 Infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. medRxiv.
14. Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, Shi J, Dai J, Cai J, Zhang T, Wu Z, He G, Huang Y (2020) Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Compur Mater Contin 63(1):537–551.
15. Description of Logistic Regression Algorithm. https://machinelearningmastery.com/logistic-regression-for-machine-learning/. [Accessed on 10 Apr 2020]

16. Description of Multinomial Naive Bayes Algorithm https://www.3pillarglobal.com/insights/document-classification-using-multinomial-naive-bayes-classifier. [Accessed on 10 Apr 2020]

17. Description of Decision Tree Algorithm: https://dataspirant.com/2017/01/30/how_decision_tree_algorithm_works/. [Accessed on 15 Apr 2020]

18. Description of Boosting Algorithm: https://towardsdatascience.com/boosting. [Accessed on 10 Apr 2020]

19. Description of Adaboost Algorithm: https://towardsdatascience.com/boosting-algorithm-adaboost-b673719ee60c. [Accessed on 10 July 2020]

20. Katuwal R, Suganthan PN (2018) Enhancing Multi-Class Classification of Random Forest using Random Vector Functional Neural Network and Oblique Decision Surfaces, Arxiv:1802.01240v1

21. Friedman JH (2002) Stochastic gradient boosting. Comput. Stat. Data Anal. 38(4):367–378.

22. Sharma, L. (Ed.), Garg, P. (Ed.). (2020). From Visual Surveillance to Internet of Things. New York: Chapman and Hall/CRC, https://doi.org/10.1201/9780429297922

23. Lavanya Sharma, Nirvikar Lohan, “Performance analysis of moving object detection using BGS techniques in visual surveillance”, in International Journal of Spatiotemporal Data Science, Inderscience, vol.1, pp. 22-53, January 2019

24. Anubhav Kumar, Gaurav Jha, Lavanya Sharma, “Challenges, Potential & future of IOT integrated with block chain”; in International Journal of Recent Technology and Engineering, Volume-8, Issue-2S7, pp.530-536,July 2019

25. Akshit Anand, Vikrant Jha, Lavanya Sharma, “An improved local binary patterns histograms techniques for face recognition for real time application”, in International Journal of Recent Technology and Engineering, Volume-8, Issue-2S7, pp. 524–529, July 2019

26. Lavanya Sharma, Dileep Kumar Yadav, “Histogram based Adaptive Learning Rate for Background Modelling and Moving Object Detection in Video Surveillance”, International Journal of Telemedicine and Clinical Practices, Inderscience, June, 2016 (ISSN: 2052-8442, DOI: 10.1504/IJTMCP.2017.082107).

27. Lavanya Sharma, Nirvikar Lohan, “Performance Analysis of Moving Object Detection using BGS Techniques”, International Journal of Spatio-Temporal Data Science, Inderscience, February, 2019.

28. Lavanya Sharma, “Introduction: From Visual Surveillance to Internet of Things”, From Visual Surveillance to Internet of Things”, Taylor & Francis, CRC Press, Vol.1, pp.14.

29. Lavanya Sharma, P K Garg “Block based Adaptive Learning Rate for Moving Person Detection in Video Surveillance”, From Visual Surveillance to Internet of Things, Taylor & Francis, CRC Press, Vol.1, pp.201.

30. Lavanya Sharma, P K Garg “Smart E-healthcare with Internet of Things: Current Trends Challenges, Solutions and Technologies”, From Visual Surveillance to Internet of Things, Taylor & Francis, CRC Press, Vol.1, pp. 215.

31. Lavanya Sharma, P K Garg, Naman Agarwal “A foresight on e-healthcare Trailblazers”, From Visual Surveillance to Internet of Things”, Taylor & Francis, CRC Press, Vol.1, pp. 235.

32. Lavanya Sharma, P K Garg “Future of Internet of Things”, From Visual Surveillance to Internet of Things, Taylor & Francis, CRC Press, Vol.1, pp. 245.

33. Lavanya Sharma, P K Garg “IoT and its applications”, From Visual Surveillance to Internet of Things, Taylor & Francis, CRC Press, Vol.1, pp.29.

34. Sharma, L. (Ed.). (2021). Towards Smart World. New York: Chapman and Hall/CRC, https://doi.org/10.1201/978100305675