ANALYZING THE DEATH RATIO OF COVID PATIENTS USING MULTIPLE LOGISTIC REGRESSION FOR OBTAINING ACCURACY

Abstract:

The outbreak of the Novel Coronavirus or the COVID-19 in various parts of the world has affected the world as a whole and caused millions of deaths. This remains an ominous warning to public health and will be marked as one of the greatest pandemics in world history. This paper aims to provide a better understanding of how various Machine Learning models can be implemented in real-world situations. Apart from the analysis done on the world figures, this paper also analyzes the current trend or pattern of Covid-19 transmission in India. With the help of datasets from the Ministry of Health and Family Welfare of India, this study puts forward various trends and patterns experienced in different parts of the world. The data to be studied has been obtained for 154 days i.e. from January 22, 2020, till June 24, 2020. For future references, the data can be further analyzed, and more results can be obtained.

CHAPTER 1

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 applications, natural language processing (NLP), intelligent robots, gaming, climate modeling, 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. One of the most significant areas of ML is forecasting, 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. There are lots of studies performed for the prediction of different diseases using machine learning techniques such as coronary artery disease, cardiovascular disease prediction, and breast cancer prediction. In particular, the study is focused on live forecasting of COVID-19 confirmed cases and study 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. 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. 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.

1.1 PROPOSED ALGORITHMS

1.1.1 POLYNOMIAL REGRESSION

- Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

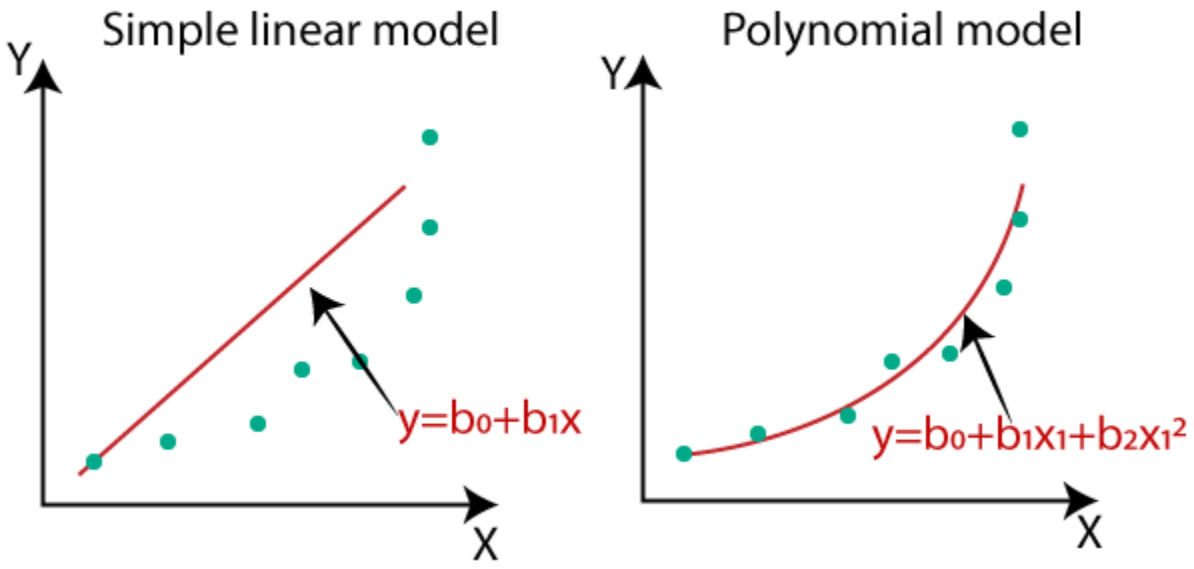
$$y= b_0+b_1x_1+ b_2x_1^2+ b_2x_1^3+..... b_nx_1^n$$

- It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.
- It is a linear model with some modification in order to increase the accuracy.
- The dataset used in Polynomial regression for training is of non-linear nature.
- It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.
- **Hence,** ***"In Polynomial regression, the original features are converted into Polynomial features of required degree (2,3,..,n) and then modeled using a linear model."***

Need for Polynomial Regression:

The need of Polynomial Regression in ML can be understood in the below points:

- If we apply a linear model on a **linear dataset**, then it provides us a good result as we have seen in Simple Linear Regression, but if we apply the same model without any modification on a **non-linear dataset**, then it will produce a drastic output. Due to which loss function will increase, the error rate will be high, and accuracy will be decreased.
- So for such cases, **where data points are arranged in a non-linear fashion, we need the Polynomial Regression model**. We can understand it in a better way using the below comparison diagram of the linear dataset and non-linear dataset.



- In the above image, we have taken a dataset which is arranged non-linearly. So if we try to cover it with a linear model, then we can clearly see that it hardly covers any data point. On the other hand, a curve is suitable to cover most of the data points, which is of the Polynomial model.

- Hence, *if the datasets are arranged in a non-linear fashion, then we should use the Polynomial Regression model instead of Simple Linear Regression.*

> ***Note:*** *A Polynomial Regression algorithm is also called Polynomial Linear Regression because it does not depend on the variables, instead, it depends on the coefficients, which are arranged in a linear fashion.*

Equation of the Polynomial Regression Model:

Simple Linear Regression equation: $y = b_0+b_1x$ **.....(a)**

Multiple Linear Regression equation: $y= b_0+b_1x+ b_2x_2+ b_3x_3+....+ b_nx_n$ **.....(b)**

Polynomial Regression equation: $y= b_0+b_1x + b_2x^2+ b_3x^3+....+ b_nx^n$ **.....(c)**

2

When we compare the above three equations, we can clearly see that all three equations are Polynomial equations but differ by the degree of variables. The Simple and Multiple Linear equations are also Polynomial equations with a single degree, and the Polynomial regression equation is Linear equation with the nth degree. So if we add a degree to our linear equations, then it will be converted into Polynomial Linear equations.

CHAPTER 2

LITERATURE REVIEW

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

The major part of the project development sector considers and fully survey all the required needs for developing the project. For every project Literature survey is the most important sector in software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations.

[2.1] COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms

The catastrophic outbreak of Severe Acute Respiratory Syndrome - Coronavirus (SARS-CoV-2) also known as COVID-2019 has brought the worldwide threat to the living society. The whole world is putting incredible efforts to fight against the

spread of this deadly disease in terms of infrastructure, finance, data sources, protective gears, life-risk treatments and several other resources. The artificial intelligence researchers are focusing their expertise knowledge to develop mathematical models for analyzing this epidemic situation using nationwide shared data. To contribute towards the well-being of living society, this article proposes to utilize the machine learning and deep learning models with the aim for understanding its everyday exponential behaviour along with the prediction of future reachability of the COVID-2019 across the nations by utilizing the real-time information from the Johns Hopkins dashboard.

[2.2] COVID-19 in India: State-wise Analysis and Prediction

The highly infectious coronavirus disease (COVID-19) was first detected in Wuhan, China in December 2019 and subsequently spread to 212 countries and territories around the world, infecting millions of people. In India, a large country of about 1.3 billion people, the disease was first detected on January 30, 2020, in a student returning from Wuhan. The total number of confirmed infections in India as of May 3, 2020, is more than 37,000 and is currently growing fast.

[2.3] COVID-19 Outbreak Prediction with Machine Learning

Several outbreak prediction models for COVID-19 are being used by officials around the world to make informed-decisions and enforce relevant control measures. Among the standard models for COVID-19 global pandemic prediction, simple epidemiological and statistical models have received more attention by

authorities, and they are popular in the media. Due to a high level of uncertainty and lack of essential data, standard models have shown low accuracy for long-term prediction. Although the literature includes several attempts to address this issue, the essential generalization and robustness abilities of existing models needs to be improved. This paper presents a comparative analysis of machine learning and soft computing models to predict the COVID-19 outbreak. Among a wide range of machine learning models investigated, two models showed promising results (i.e., multi-layered perceptron, MLP, and adaptive network-based fuzzy inference system, ANFIS). Based on the results reported here, and due to the highly complex nature of the COVID-19 outbreak and variation in its behavior from nation-to-nation, this study suggests machine learning as an effective tool to model the outbreak.

[2.4] Forecasting the novel coronavirus COVID-19

What will be the global impact of the novel coronavirus (COVID-19)? Answering this question requires accurate forecasting the spread of confirmed cases as well as analysis of the number of deaths and recoveries. Forecasting, however, requires ample historical data. At the same time, no prediction is certain as the future rarely repeats itself in the same way as the past. Moreover, forecasts are influenced by the reliability of the data, vested interests, and what variables are being predicted. Also, psychological factors play a significant role in how people perceive and react to the danger from the disease and the fear that it may affect them personally. This paper introduces an objective approach to predicting the continuation of the COVID-19 using a simple, but powerful method to do so. Assuming that the data used is reliable and that the future will continue to follow the past pattern of the disease, our forecasts suggest a continuing increase in the confirmed COVID-19

cases with sizable associated uncertainty. The risks are far from symmetric as underestimating its spread like a pandemic and not doing enough to contain it is much more severe than overspending and being over careful when it will not be needed. This paper describes the timeline of a live forecasting exercise with massive potential implications for planning and decision making and provides objective forecasts for the confirmed cases of COVID-19.

[2.5] Prediction of Spreads of COVID-19 in India from Current Trend

The article describe modelling efforts for evaluating the current level of COVID-19 infections in India, using exponential model. The Data from 15 march 2020 to 30 April 2020 are used for validating the model, where intrinsic rise rate is kept constant. It is observed that some states of India, like MAharastra, Gujarat and Delhi have a much higher daily infection cases. This is modelled by assuming an initial higher infections, keeping rise rate same. The sudden outbursts are captured using offset of values for these three states. Data from other states like Madhya Pradesh, Uttar Pradesh and Rajasthan are also analysed and they are found to be following the same constants as India is following. Worldwide, many attempts are made to predict outburst of COVID-19 and in the model, described in this paper, turning point is not predicted, as cases in India are still rising. The developed model is based on daily confirmed infections and not on cumulative infections and rationalization is carried out for the population of various regions, while predicting infections for various states. Assigning a decay constant at this stage will be a premature exercise and keeping that in mind, exponential model predicts that India will attain 1 lakh case by 15 May 2020. The figure of 2 lakh and 3 lakh will be attained on 22 May 2020 and 26 May 2020, respectively.

[2.6] An interpretable mortality prediction model for COVID-19 patients

The sudden increase in COVID-19 cases is putting high pressure on healthcare services worldwide. At this stage, fast, accurate and early clinical assessment of the disease severity is vital. To support decision making and logistical planning in healthcare systems, this study leverages a database of blood samples from 485 infected patients in the region of Wuhan, China, to identify crucial predictive biomarkers of disease mortality. For this purpose, machine learning tools selected three biomarkers that predict the mortality of individual patients more than 10 days in advance with more than 90% accuracy: lactic dehydrogenase (LDH), lymphocyte and high-sensitivity C-reactive protein (hs-CRP). In particular, relatively high levels of LDH alone seem to play a crucial role in distinguishing the vast majority of cases that require immediate medical attention. This finding is consistent with current medical knowledge that high LDH levels are associated with tissue breakdown occurring in various diseases, including pulmonary disorders such as pneumonia. Overall, this Article suggests a simple and operable decision rule to quickly predict patients at the highest risk, allowing them to be prioritized and potentially reducing the mortality rate.

CHAPTER 3

SYSTEM REQUIREMENTS

3.1 HARDWARE REQUIREMENTS

System : Pentium Dual Core.

Hard Disk : 120 GB.

Monitor : 15”LED

Input Devices : Keyboard, Mouse

Ram : 1GB.

3.2 SOFTWARE REQUIREMENTS

Operating system : Windows 7.

Coding Language : python

Toolkit : Jupitor Notebook

DATABASE : EXCEL