PREDICTION OF COVID-19 USING MACHINE LEARNING APPROACHES

UTKARSH ANIL BAVISKAR

VISHWAKARMA INSTITUTE OF INORMATION TECHNOLOGY, PUNE

Abstract - *The outbreak of the COVID19 virus, called SARSCoV2, has created a pandemic situation worldwide. The cases of COVID19 are increasing rapidly every day. Machine learning (ML) and cloud computing can be implemented very effectively to track diseases, predict epidemic growth, and develop strategies and guidelines to control their spread. This study uses improved mathematical models to analyze and predict the growth of virus. We applied an improved ML-based model to predict potential COVID-19 threats in countries around the world. Prediction of COVID-19 can be done by iteratively weighting to approximate the generalized inverse Weibull distribution. It has been deployed on cloud platforms to more accurately and realistically predict the dynamics of epidemic growth. It is a more accurate data-driven approach as it can be very useful for government and citizens of the nation. Hence, we propose a research and setup grounds for further research.*

Key Words: **Machine learning, COVID-19, AI, SVM, Random Forest, Decision Tree, Linear Regression**

1. INTRODUCTION

The novel coronavirus infection (COVID-19) was first reported in Wuhan, Hubei Province, China on December 31, 2019. It began to spread rapidly around the world. The cumulative incidence of this virus (SARSCoV2) is increasing rapidly and has affected 196 countries and territories, of which the United States, Spain, Italy, United Kingdom and France have been most affected. WHO has declared a global pandemic for the coronavirus infection (COVID-19), and the virus continues to spread. As of May 4, 2020, there were 3,581,884 confirmed cases and 248,558 deaths. The main difference between the CoV2 pandemic and related viruses such as severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) is that CoV2 spreads rapidly through human contact and nearly 20% of infected subjects remain asymptomatic carrier. In addition, various studies have reported that CoV2-induced disease is more at risk for people with weakened immune systems. The elderly and those with life-threatening diseases such as cancer, diabetes, neurological diseases, coronary heart disease and HIV/AIDS are more vulnerable to the serious consequences of COVID-19. In the absence of drugs, the only solution is to slow the spread of the virus by applying "social distancing" to break the transmission chain. This behavior of CoV2 requires the development of a robust mathematical framework to track the spread and the automation of tracking tools to make dynamic decisions online. Innovative solutions are needed to develop, manage, and analyze big data for growing target networks, patient information, and big data for movement within communities, as well as integration with clinical trial and pharmaceutical data, genomic data and public health data. Multiple data sources, including text messages, online communications, social media, and web articles, can be very useful in analyzing the increase in infections caused by community behavior. By wrapping this data with machine learning (ML) and artificial intelligence (AI), researchers can predict when and where a disease may spread and notify the area to agree on the necessary action. By automatically tracking the travel history of infected subjects, you can study epidemiologic correlations with disease spread in specific communities.

2. MOTIVATION AND OUR CONTRIBUTIONS

ML can be used to process large amounts of data and intelligently predict the spread of a disease. Cloud computing can be used to rapidly improve the forecasting process with high-speed computing. New energy-efficient peripheral systems can be used to collect data to reduce power consumption. In this article, we present a predictive model deployed using the FogBus framework to accurately predict the number of COVID-19 cases, an increase and decrease in the number of cases in the near future, and the date the pandemic can be expected to end in other countries. We also provide a detailed comparison with the baseline model and show how devastating the impact can be if a poorly matched model is used. We empower governments and citizens to be proactive by presenting a prediction framework based on machine learning models that can be used to make real-time predictions from remote cloud nodes. In conclusion, we summarize this work and present various lines of research.

3. SOFTWARE PLATFORMS

- Python

Python is an open source programming language. Currently in high demand in the IT industry. It is mainly used for machine learning for website development, data processing, software, etc. Python is almost similar to C, except that the coding syntax is different. You can perform many types of tasks and use them to build machine learning, data analysis, and complex statistical calculations.

- Machine Learning Model

Many recent studies have shown that the spread of COVID-19 follows an exponential distribution. Empirical estimates of the SARSCoV2 pandemic and previous data sets have shown that many sources have a large number of outliers in the data corresponding to new cases over time, which may or may not follow a standard distribution such as a Gaussian or exponential distribution. In a recent study by the Singapore University of Technology and Design (SUTD) Data-Driven Innovation Lab, a Susceptible Infected Recovered model was used to construct a regression curve and distributed a Gaussian distribution to estimate the number of cases over time. However, in the previously described study, an older version of the virus called SARSCoV2 followed the generalized inverse Weibull distribution (GIW) better than Gaussian.

4. PREDICTION MODEL AND PERFORMANCE COMPARISONS

The machine learning (ML) and data science communities are working hard to improve the predictions of epidemiologic models and analyze information sent via Twitter to develop governance strategies and evaluate the impact of policies to contain the spread. Various data sets have been published publicly on this topic. However, as COVID-19 spreads globally, more data needs to be collected, processed and analyzed. The novel coronavirus is having a serious socio-economic impact worldwide. Countries with large populations should be more vigilant as the virus is easily transmitted through droplets or runny nose, mainly when an infected person coughs or sneezes. To get detailed information about the impact of COVID-19 on the world's population, and to predict the number of COVID-19 cases in different countries and when the pandemic is expected to end, we propose a machine learning model that can be run continuously in cloud data centers. (CDC). Accurately predict outbreaks and proactively develop strategic responses.

5. DATASETS

The dataset used in this case study is our world in Data2 by Hannah Ritchie. The dataset is updated daily based on status reports from the World Health Organization (WHO). More information about the dataset can be found on our website: https://ourworldindata.org/coronavirus-source-data.

6. ALGORITHMS

1. SVM

SVM was chosen because it transforms an inseparable problem into a separable problem by using a kernel trick to transform a low-dimensional input space into a high-dimensional space. We split the dataset into a train set and a test set in a ratio of 7:3, and using a linear kernel, the SVM classifier uses a hyperplane to linearly divide the data. Each data class is separated by parallel hyperplanes to keep distances as large as possible.

2. Random Forest,

Random Forest is an ensemble technique that can perform both regression and classification tasks using multiple decision trees and techniques commonly known as bootstrap and aggregation known as packaging. The main idea behind this is to combine multiple decision trees to determine the final result instead of relying on separate decision trees.

The Random Forest has several decision trees as the basic learning model. We perform row sampling randomization and feature sampling from the dataset forming a sample dataset for each model. This section is called Bootstrap.

3. Decision Tree

Decision trees are very successful classifiers applied in many domains. Decision trees are constructed using a recursive partitioning process in which data points are partitioned at each node using selected partitioning criteria. The path from the root node to the sheet is the rule used for prediction. An ensemble of classifiers consists of a set of classifiers [18]. The final decision is the combination of all member classifiers. Ensembles generally perform better than individual members when their individual members are precise and varied. Decision tree ensembles are fairly robust and perform well. The experiment uses several ensembles of decision trees. Decision tree ensembles designed for unbalanced datasets are also used because the data are unbalanced.

4. Linear regression

Since we are dealing with COVID-19 data, observations suggest that COVID-19 data do not follow a linear property. Rather, it follows a linear property for a short time and then changes direction. In this case, it is not suitable for linear regression, and if you still use it for linear regression, the prediction is far from real.

6. MAE/MSE SCORE

Name of Algorithm	MAE	MSE
SVM	113952693.00743757	1.4189100362824158e+16
Linear Regression	29503683.939641554	876863594001211.8
Decision Tree	25613676.38095238	869104703463814.2
Random Forest Classifier	25613676.38095238	869104703463814.2

7. RESULTS

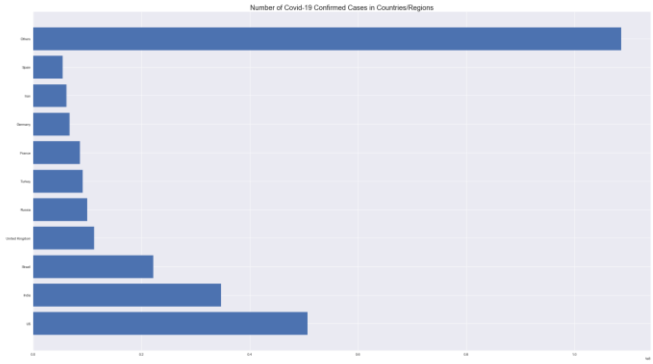


Fig:7.1: Number of Cases

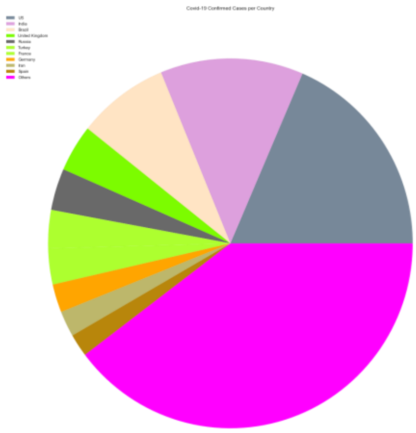


Fig.7.2: Number of Cases pie-chart

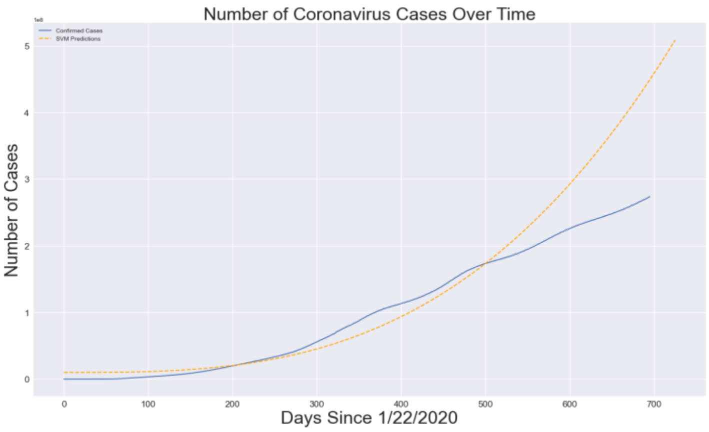


Fig.7.3: SVM Predictions

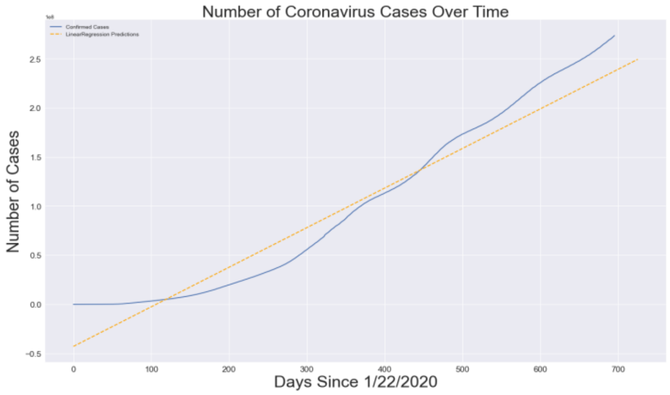


Fig.7.4: Linear Regression Prediction

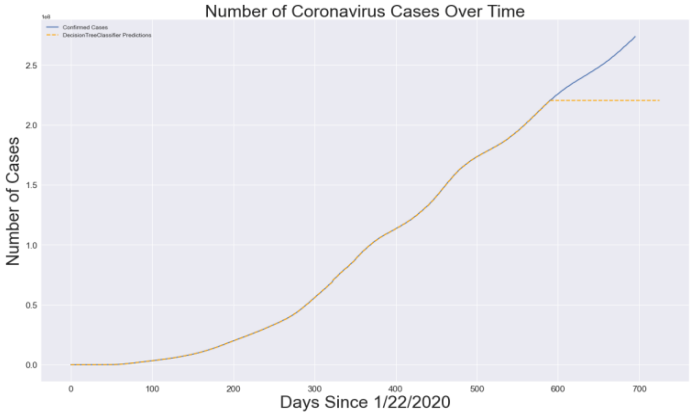


fig.7.5: Decision Tree Prediction

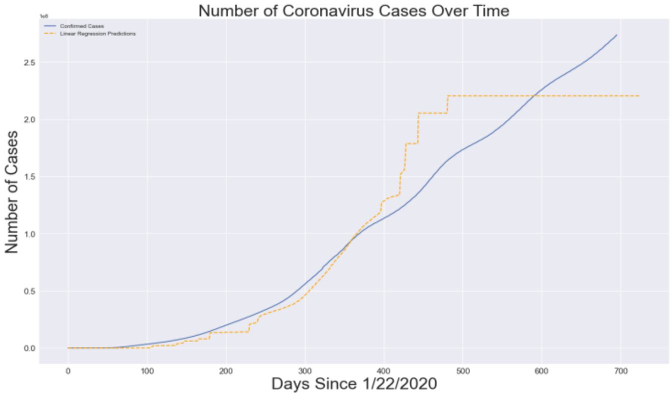


fig.7.6: Random Forest Prediction

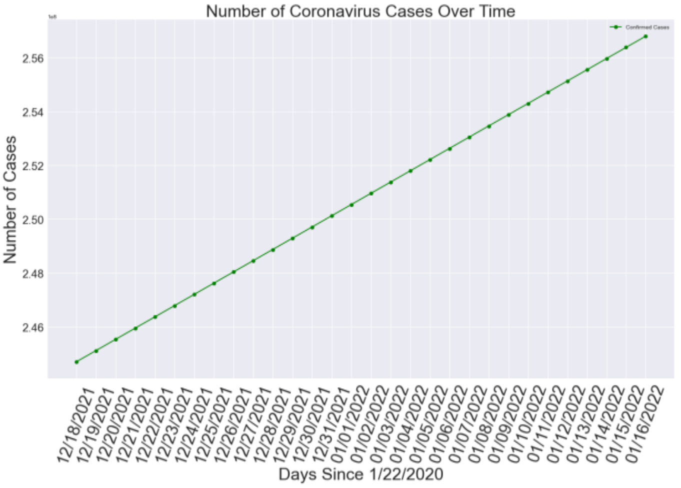
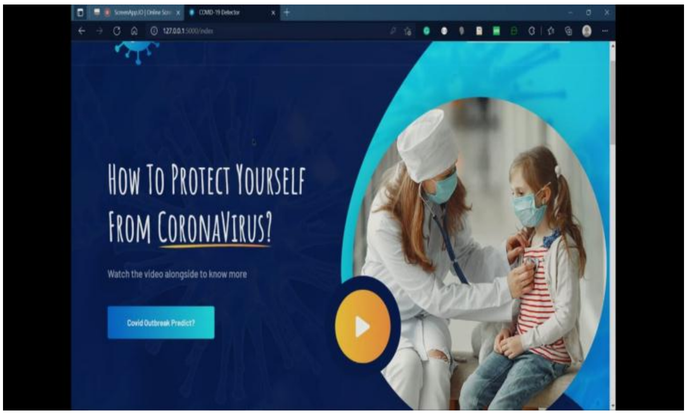


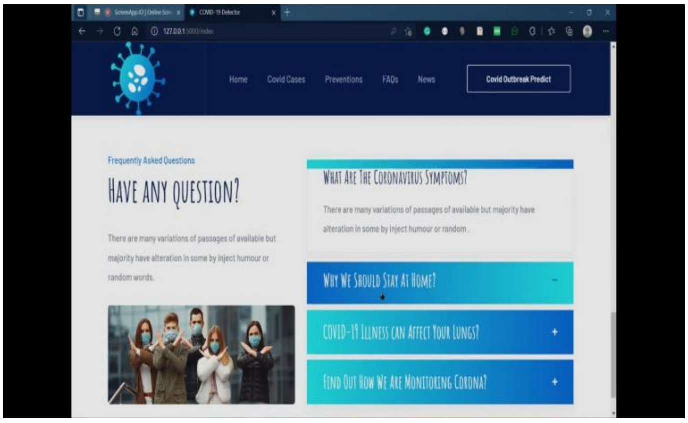
fig.7.7: Final Prediction

8. OUTPUT SCREENS


How To Protect Yourself From Coronavirus?

Watch the video alongside to know more

fig.8.1: Dashboard



Frequently Asked Questions

Have any question?

There are many variations of passages of available but majority have alteration in some by inject humour or random words.

What Are The Coronavirus Symptoms?

There are many variations of passages of available but majority have alteration in some by inject humour or random .

Why We Should Stay At Home?

COVID-19 Illness can Affect Your Lungs?

Find Out How We Are Monitoring Corona?

Fig.8.2: Dashboard with FAQ questions

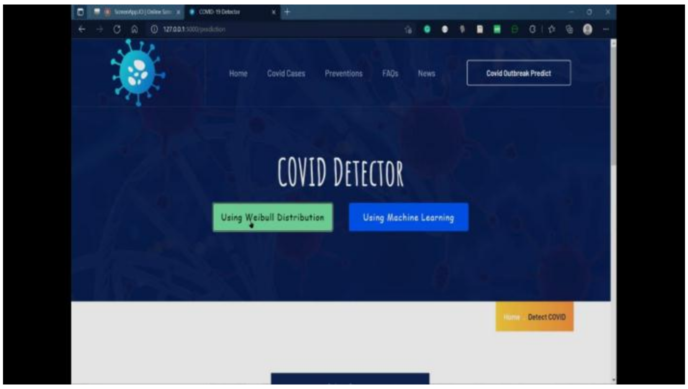


Fig.8.3: Prediction using Weibull Distribution

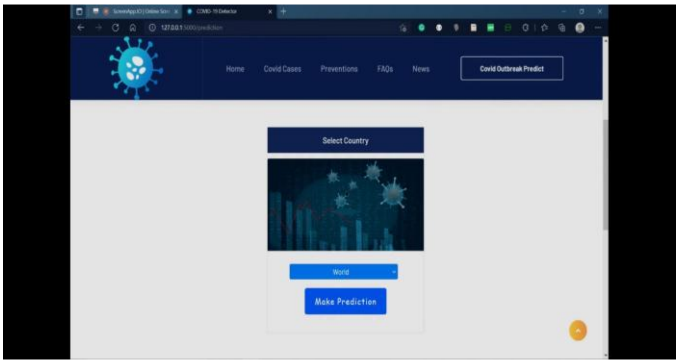


Fig.8.4. Select country for prediction with Weilbull Distribution

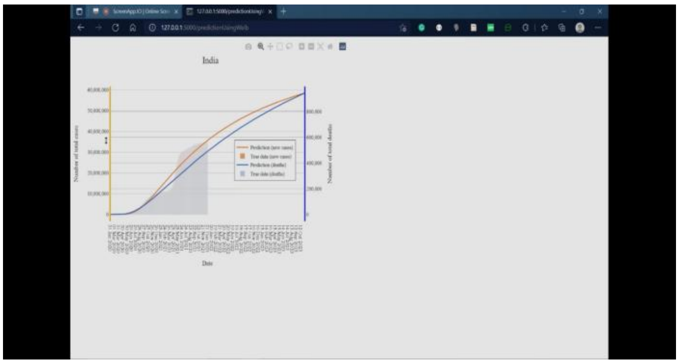


Fig.8.5. Final Prediction using Weilbull Distribution

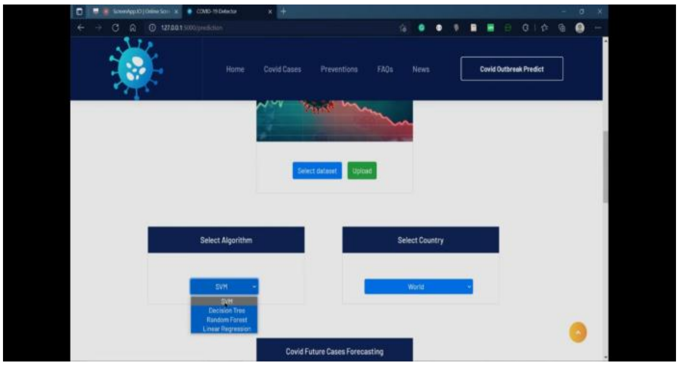


fig.8.6: Prediction using ML algorithms

9. SUMMARY AND CONCLUSIONS

In this study, we discussed how machine learning, and cloud computing can help in predicting of the growth of pandemic. Additionally, case studies have been published demonstrating the severity of the spread of CoV-2 in countries around the world. Using the proposed robust Weibull model based on iterative weighting, we show that our model can make statistically better predictions than the baseline. The baseline Gaussian model shows an overly optimistic picture of the COVID-19 scenario. SVM algorithm having MAE as 113952693.00743757 and MSE as 1.4189100362824158e+16.

10. REFERENCES

[1] COVID Live - Coronavirus Statistics - Worldometer. (2021).https://www.worldometers.info/coronavirus/R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[2] Wang, C., Horby, P. W., Hayden, F. G., & Gao, G. F. (2020). A novel coronavirus outbreak of global health concern. *The Lancet*, *395*(10223), 470–473. https://doi.org/10.1016/s0140-6736(20)30185-9.

[3] Guangdi Li and Erik De Clercq. Therapeutic options for the 2019 novel coronavirus (2019-ncov), 2020.

[4] Smriti Mallapaty. What the cruise-ship outbreaks reveal about covid-19. Nature, 580(7801):18–18, 2020.

[5] Kai Liu, Ying Chen, Ruzheng Lin, and Kunyuan Han. Clinical features of covid-19 in elderly patients: A comparison with young and middle-aged patients. Journal of Infection, 2020.

[6] Shi Zhao, Qianyin Lin, Jinjun Ran, Salihu S Musa, Guangpu Yang, Weiming Wang, Yijun Lou, Daozhou Gao, Lin Yang, Daihai He, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. International Journal of Infectious Diseases, 92:214–217, 2020.

[7] Shreshth Tuli, Shikhar Tuli, Gurleen Wander, Praneet Wander, Sukhpal Singh Gill, Schahram Dustdar, Rizos Sakellariou, and Omer Rana. Next generation technologies for smart healthcare: Challenges, vision, model, trends and future directions. Internet Technology Letters, page e145.

[8] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet, 395(10223):497–506, 2020.