Analysis of Covid-19 (India) Using Machine Learning Algorithms

Prof. S. B. Nikam[1], Tanish Jain[2], Akhil Aditya[3], Yashraj Tandon[4]

[1]*Assistant Professor, Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering Pune*

[2,3,4]*Student, Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering Pune*

***Abstract* - In light of recent events, such as the coronavirus pandemic, prediction algorithms based on machine learning (ML) have proven effective in predicting perioperative outcomes and improving decision-making in the future. Machine learning models have long been used in many application areas that need to detect and prioritize negative threat characteristics. Typically, a variety of forecasting methods are used to address forecasting problems. This study shows how machine learning algorithms can predict the number of patients who will be infected by COVID19, a virus that is now considered a possible threat to humans. In this study, the following predictive model are used to predict COVID19 risk factors: Linear Regression, Exponential Time Smoothing, Autoregressive Integrated Moving Average (ARIMA). The results of the study indicate that these strategies are a viable option in the current COVID19 pandemic.**

Index Terms - **COVID, Machine Learning**

INTRODUCTION

This research aims to provide an early prediction model for the spread of the new coronavirus (also known as SARS CoV2, officially called COVID19 by the World Health Organization (WHO)) [1]. COVID19 is currently a very serious threat to human life around the world. In late 2019, the virus was first discovered in Wuhan, China, when large numbers of people developed symptoms such as pneumonia [2]. It has many effects on the human body, including severe acute respiratory syndrome and multiple organ failure, which will eventually lead to death in a short period of time [3]. Hundreds of thousands of people around the world are affected by this epidemic and thousands of people will die every day in the future. According to reports, thousands of new people in countries around the world test positive every day. The virus is mainly spread through close physical contact between people, respiratory droplets, or contact with contaminated surfaces. The most challenging aspect of its spread is that a person can be infected with the virus for many days without symptoms. Taking into account the causes and dangers of its spread, almost all countries have announced partial or strict blockades of affected areas and cities. At present, medical researchers all over the world are working to find suitable vaccines and drugs for the disease. Since there are no approved drugs that can kill the virus, governments in all countries are paying attention to preventive measures that can stop the spread. Among all preventive measures, “knowing” all aspects of COVID19 is considered extremely important. To provide this information, many researchers are studying different aspects of the pandemic and producing results that are helpful to humans.

By solving many very complex and complex real-world problems, machine learning (ML) has proven to be a prominent research area in the past decade. Application areas include almost all real-world areas, such as healthcare, autonomous vehicles (AV), commercial applications, natural language processing (NLP), intelligent robots, games, weather modelling, voice and image processing. Learning ML algorithms is usually based on trial and error, which is the exact opposite of traditional algorithms, which follow programming instructions based on decision statements such as if else. One of the most important areas of ML is prediction. Many standard ML algorithms have been used in this area to guide the process of taking necessary actions in many application areas in the future, including weather forecasting, disease prediction, stock market prediction, and disease. forecast. Several regression and neural network models have wide applicability and can be used to predict the condition of patients with specific diseases in the future. There are many studies using machine learning techniques to predict different diseases, such as coronary artery disease, cardiovascular disease prediction [4], and breast cancer prediction. In particular, research is focused on the prognosis of confirmed COVID19 cases, and research also focused on the prognosis of COVID19 outbreaks and early responses. These predictive systems are very useful in decision-making, can manage the current situation and guide early intervention to manage these diseases very effectively.

II.RELATED WORK

1. 'Machine Learning to Predict COVID-19 and ICU Requirement' - This paper focuses on the application of machine learning (ML) algorithms to manage novel coronavirus disease (COVID-19) Different ML classifiers are used for two cases, one for the prediction of covid 19 patients and another for ICU requirement [5].

2. 'Regression Analysis of COVID-19 using Machine Learning Algorithms' - The outbreak of the Novel Coronavirus or the COVID-19 has affected the world as a whole and caused millions of deaths. This document aims to better understand how to implement various machine learning models in the real world. The data to be studied has been obtained for 154 days i.e., from January 22, 2020, till June 24, 2020 [6].

3. 'ANN based COVID -19 Prediction and Symptoms Relevance Survey and Analysis'
- The main focus of 0this Research is to see
how one can easily predict if he/she has been infected by COVID-19. The virus has affected more than 66,729,375 people across 220 countries and has also cost the lives of 1,535,982 people. Research still predicts that another second wave is to hit soon [7].

4. 'COVID-19 Future Forecasting Using Supervised Machine Learning Models' - Machine learning (ML) based forecasting mechanisms have proved their significance to anticipate in perioperative outcomes. This research demonstrates the ability of the ML model to predict the number of next patients affected by COVID19. The results produced by the study proves it a promising mechanism to use these methods for the current scenario of the pandemic [8].

5. 'Covid-19 Outbreak Modelling Using Regression Techniques' - Since the onset of COVID-19 pandemic, officials and several others have been making an effort to form informed decisions and take relevant measures to curb the outbreak. They use standard statistical models and epidemiological models to determine the spread of the epidemic.
Although these models have shown to have accuracy in the past, they seem to be highly ineffective during the current outbreak [9].

6. 'Machine learning models for covid-19 future forecasting' - Machine learning algorithms have been applied for a long time in many applications requiring the detection of adverse risk factors. This study shows the ability of modelling ML to predict the number of people affected by COVID19 as a potential threat to humans [10].

7. 'A Comparative Approach to Predict Corona Virus Using Machine Learning' - More than 250 countries have been affected by Coronavirus disease (COVID-19) The Indian government is making the necessary steps to control the spread of virus in the society. This paper presents the prediction and analysis of the possible threat of Covid-19 using various machine learning algorithms. It uses 20 metrics including the patient's geographical location, travel history, health record statistics, etc., to predict the severity of the case and the feasible outcome [11].

8. 'Predicting COVID-19 infections and deaths in Bangladesh using Machine Learning Algorithms' - Bangladesh is in the midst of the community spread of the new coronavirus (COVID19). Since December 2019, more than 700,000 people have died and more than 10 million people have been infected. Study has shown that when predicting pandemic situations, the Facebook Prophet model provides the highest accuracy among many predictive models. [12].

9. 'The Prediction of the Spread of COVID-19 using Regression Models' - The COVID-19 outbreak has spread to over 200 nations, making it an unparalleled public health problem that has a significant impact on people's daily lives. Machine learning and a variety of regression models, including linear regression, polynomial regression, multi-regression, and Lasso regression models, were used in this study [13].

10. 'Predicting the Existence of COVID-19 using Machine Learning Based on Laboratory Findings' - A novel coronavirus illness (COVID-19) has been

identified in Wuhan, China, and has since spread around the world. We emphasis the illness prediction process using a mix of wrapper feature selection (FS) and four distinct classifiers in this paper [14].

III. TECHNIQUE

For visualization and analysis, the following software's were used: -

1. Tableau – Tableau software is one of the fastest growing data visualization tools used in the BI industry today. This is the best way to change or convert the original data set into an easy-to-understand format with zero technical skills and coding knowledge. Tableau is widely used because it can analyze data very quickly. In addition, visualizations are generated as dashboards and worksheets. Tableau allows you to create dashboards to provide useful information and drive business development. If Tableau products are configured with the correct underlying hardware and operating system, they will always run in a virtualized environment. Tableau is used to explore data through unlimited visual analysis [15].
2. Alteryx - Alteryx's core product is its workflow designer. Alteryx Designer enables users to quickly prepare, mix, integrate, and analyze data from almost any source, such as flat files, database connections, APIs, salesforce.com, and more. All of this is done using a visual workflow designer, which is very intuitive and requires no code. Workflows are created using a combination of predefined or custom "tools". The tools represent functional units, such as data input, data output, aggregation, filtering, sampling, reporting, geocoding, and so on. Although workflows can become very large, they are essentially just a series of connected tools. A particularly cool feature of Alteryx is that any user- created workflow can be saved as a tool and reused in other workflows. Alteryx refers to these customization tools as "macros", and they allow extensive customization and reuse of your work [16].

For forecasting following approaches were used –

1. Linear Regression

The idea behind linear regression is that you can use the straight line of best fit (also called the regression line) to determine if there is a relationship (correlation) between the dependent variable (Y) and the independent variable (X). Equation is Y = a + bX. Y is the dependent variable, a is the y- intercept, b is the slope of the line, and X is the independent variable. You can use this equation to predict where the data points will fall based on a given predictor variable [17].

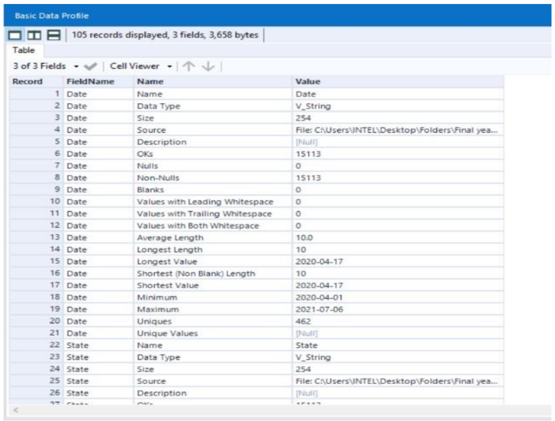
2. Exponential Time Smoothing

The exponential smoothing method is a series of forecasting models. They use the weighted average of past observations to predict new values. Here, the idea is to pay more attention to the recent value of the series. Therefore, as the observed values age (in time), the importance of these values decreases exponentially. The exponential smoothing method combines error, trend, and seasonal components in the smoothing calculation. Each item can be added, multiplied or removed from the model. These three terms (error, trend, and season) are called ETS. Therefore, the exponential smoothing method can be defined according to the ETS framework, in which the components are calculated in a smooth manner [18].

3. Autoregressive Integrated Moving Average (ARIMA)

Autoregressive Integrated Moving Average (ARIMA) is a statistical analysis model that uses time series data to better understand a data set or predict future trends. A statistical model is autoregressive if it predicts future values based on past values. ARIMA uses lagged moving averages to smooth time series data. is widely used in technical analysis to predict the future price of securities. The autoregressive model implicitly assumes that the future will be similar to the past [19].

IV.TESTING

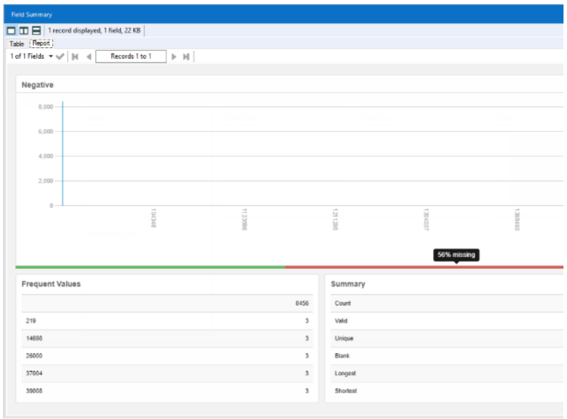


Basic Data Profile

105 records displayed, 3 fields, 3,658 bytes

Record	FieldName	Name	Value
1	Date	Name	Date
2	Date	Data Type	V_String
3	Date	Size	254
4	Date	Source	File: C:\Users\INTEL\Desktop\Folders\Final yea...
5	Date	Description	[Null]
6	Date	OKs	15113
7	Date	Nulls	0
8	Date	Non-Nulls	15113
9	Date	Blanks	0
10	Date	Values with Leading Whitespace	0
11	Date	Values with Trailing Whitespace	0
12	Date	Values with Both Whitespace	0
13	Date	Average Length	10.0
14	Date	Longest Length	10
15	Date	Longest Value	2020-04-17
16	Date	Shortest (Non Blank) Length	10
17	Date	Shortest Value	2020-04-17
18	Date	Minimum	2020-04-01
19	Date	Maximum	2021-07-06
20	Date	Uniques	462
21	Date	Unique Values	[Null]
22	State	Name	State
23	State	Data Type	V_String
24	State	Size	254
25	State	Source	File: C:\Users\INTEL\Desktop\Folders\Final yea...
26	State	Description	[Null]
27	State	OKs	15113

Fig: Basic Data Profile



Field Summary

1 record displayed, 1 field, 22 KB

Table | Report

1 of 1 Fields | Records 1 to 1

Negative

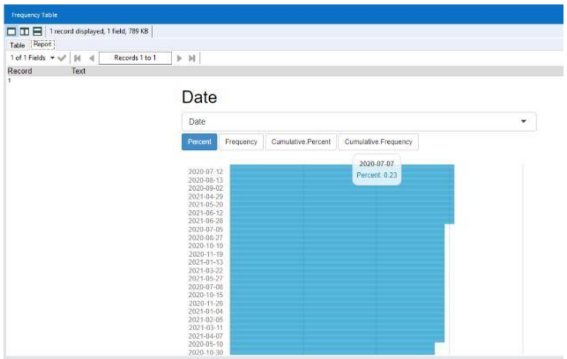
Frequent Values

	8456
219	3
14698	3
26000	3
37004	3
39008	3

Summary

Count
Valid
Unique
Blank
Longest
Shortest

Fig: Field Summary



Frequency Table

1 record displayed, 1 field, 789 KB

Table | Report

1 of 1 Fields | Records 1 to 1

Record | Text

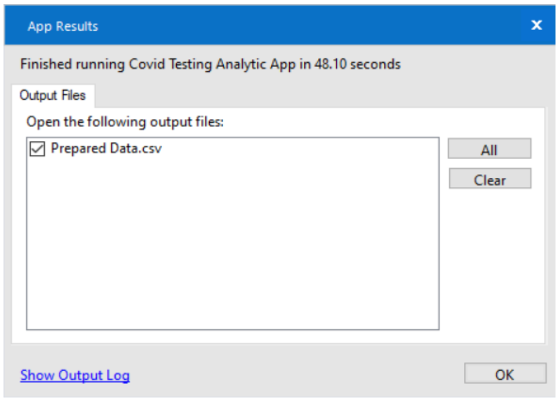
1

Date

Date

Percent | Frequency | Cumulative Percent | Cumulative Frequency

Fig: Frequency Table



App Results

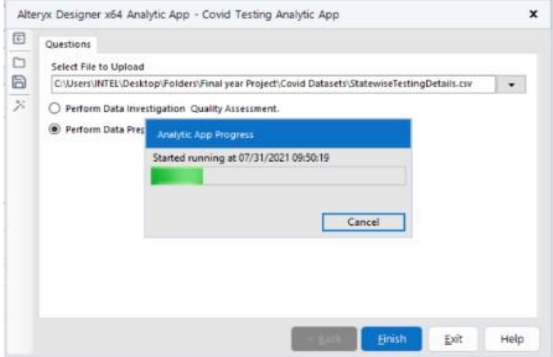
Finished running Covid Testing Analytic App in 48.10 seconds

Output Files

Open the following output files:

☑ Prepared Data.csv

Fig: App Results



Alteryx Designer x64 Analytic App - Covid Testing Analytic App

Questions

Select File to Upload

C:\Users\INTEL\Desktop\Folders\Final year Project\Covid Datasets\StatewiseTestingDetails.csv

○ Perform Data Investigation Quality Assessment.

● Perform Data Prep

Analytic App Progress

Started running at 07/31/2021 09:50:19

Fig: Running Analytics App

V.RESULTS

No of Deaths						
Initial	Change From Initial	Seasonal Effect		Contribution		
		High	Low	Trend	Season	Quality
26 May 2021	26 May 2021 – 5 June 2021	5 June 2021	1 June 2021			
1 ± 4 7	4	6	-2	2.0 %	98.0 %	Ok
		ETS	Forecast	Results		
No of Deaths						

Model		Quality Metrics						Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	Additive	Additive	18	13	0.78	195.8%	225	0.110	0.000	0.104
			E TS	forecast		Results				

Record	Report				
1	Basic Summary				
2	Call:				
3	lm(formula = Positive ~ State + Negative, data = the.data)				
4	Residuals:				
5	Min -320093.6	1Q -7611.5	Median 34.4	3Q 7613.1	Max 714813.4

Linear Model Results

Record	Report		
1	Summary of ARIMA model ARIMA_Vaccine_Forecast		
2	Method: ARIMA(0,1,0)(0,0,0)[7]		
3	Call: Auto.arima(Total.Doses.Administered)		
4	Coefficients: Value Std Err		
5	Sigma^2 estimated as 1112844537213.55: log likelihood = -94647.44328		
6	Information Criteria:		
	AIC 189296.8866	AICc 189296.8872	BIC 189303.6174
7	In-sample error measures:		

	ME 3747.608	RMSE 1054829.282	MAE 112267.785	MPE -Inf	MAPE Inf	MASE 0.1491701	ACF1 - 0.0020628
8	Ljung-Box test of the model residuals: Chi-squared = 8.1923, df = 24, p-value = 0.998883						

ARIMA Model Results

VI. CONCLUSION

As per the results of visualizations of our model, the second wave of covid-19 in India had a very deadly impact. We also got to see some seasonality in the wave pattern. During the winter season corona cases were low, whereas in summers corona cases were at a high.

REFERENCES

[1] Naming the Coronavirus Disease (Covid-19) and the Virus That Causes it, Apr. 2020

[2] 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.

[3] L. van der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. Berkhout, K. C. Wolthers, et al., "Identification of a new human Coronavirus", Nature Med., vol. 10, no. 4, pp. 368-373, 2004

[4] K. M. Anderson, P. M. Odell, P. W. Wilson and W. B. Kannel, "cardiovascular disease risk profiles", Amer. heart J., vol. 121, no. 1, pp. 293-298, 1991.

[5] P. Podder and M. R. H. Mondal, "Machine Learning to Predict COVID-19 and ICU Requirement," 2020 11th International Conference on Electrical and Computer Engineering (ICECE), 2020, pp. 483-486, doi: 10.1109/ICECE51571. 2020. 9393 123.

[6] E. Gambhir, R. Jain, A. Gupta and U. Tomer, "Regression Analysis of COVID-19 using Machine Learning Algorithms," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 65- 71, doi:10.1109/ICOSEC49089. 2020. 9215356.

[7] D. N and K. K. G, "ANN based COVID -19 Prediction and Symptoms Relevance Survey and Analysis," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1805- 1808, doi: 10.1109/ICCMC51019.2021.9418448.

[8] F. Rustam et al., "COVID-19 Future Forecasting Using Supervised Machine Learning Models," in IEEE Access, vol. 8, pp. 101489-101499, 2020, doi: 10.1109/ACCESS.2020.2997311.

[9] A. Bansal and U. Jayant, "Covid-19 Outbreak Modelling Using Regression Techniques," 2021 International Conference on Innovative Practices in Technology and Management (ICIPTM), 2021, pp. 113-118, doi: 10.1109/ICIPTM 52218. 2021. 9388347.

[10] Mojjada RK, Yadav A, Prabhu AV, Natarajan Y. Machine Learning Models for covid-19 future forecasting [published online ahead of print, 2020 Dec 9]. Mater Today Proc. 2020;10.1016 /j.matpr.2020.10.962.doi:10.1016/j. matpr. 2020. 10.962

[11] M. Rohini, K. R. Naveena, G. Jothipriya, S. Kameshwaran and M. Jagadeeswari, "A Comparative Approach to Predict Corona Virus Using Machine Learning," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp.331-337, doi: 10.1109/ICAIS50930.2021.9395827.

[12] M. I. Leon, M. I. Iqbal, S. M. Azim and K. A. Al Mamun, "Predicting COVID-19 infections and deaths in Bangladesh using Machine Learning Algorithms," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021, pp. 70-75, doi: 10.1109/ICICT4SD50815. 2021. 9396820.

[13] Z. Li, S. Yang and J. Wu, "The Prediction of the Spread of COVID-19 using Regression Models," 2020 International Conference on Public Health and Data Science (ICPHDS), 2020, pp.247-252, doi: 10.1109/ICPHDS51617.2020.00055.

[14] H. Turabieh and W. Ben Abdessalem Karaa, "Predicting the Existence of COVID-19 using Machine Learning Based on Laboratory Findings," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), 2021, pp. 1-7, doi: 10.1109/WiDSTaif52235.2021.9430233.

[15] https://intellipaat.com/blog/what-is-tableau/
[16] https://www.arkatechture.com/blog/what-is-alteryx
[17] https://thedatastudent.com/what-is-linear-regression-a-simple-explanation/
[18] https://www.daitan.com/innovation/exponentia
[19] l-smoothing-methods-for-time-series-forecasting/https://www.investopedia.com/terms/a/autoreg ressive-integrated-moving-average-arima.asp