An Efficient Covid-19 Vaccine Analysis Using Regression Algorithms for Health Care Monitoring

Lavanya.S[1], Priyanka.M. N[2], J.T. Thirukrishna [3]
[1,2] UG Scholar, Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore, India
[3]Associate Professor, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bangalore, India
Email address: [1]lavanyashankarsv09@gmail.com , [2]priyankamn0705@gmail.com , [3]dr.tkrishna-ise@dsatm.edu.in

Abstract*— The Covid-19 epidemic is the most pivotal interest knock that has circled the earth for as long as a period. Forefeeling the COVID-19 inoculation arrangement has go a catchy issue. The ascent in different immunizations created by consummate investigators prodded interest group in addressing familiar with advancing antibody methodologies. To bring this affliction to an end, a large share of the world needs to be holy to the pesticide. The harmless way to achieve this is with a vaccine. Vaccines are a technology that humanity has hourly counted on in the history to bring down the death rate of epidemic complaints. Within lower than 12 months after the threshold of the COVID-19 illness, several examination centres constructed up to the group and developed vaccines that keep from SARS-CoV-2, the virus that causes COVID-19. Now the challenge is to make these vaccines attainable to people around the globe. It'll be critical that people all around the world enter the claimed protection. We have used linear, multilinear, and polynomial regression to predict the number of people vaccinated. We have used regression because data is linearly correlated. We have used evaluation parameters like r2 score and MAE to evaluate the performance. Through this we predict the number of people getting vaccinated in the future.

Keywords*— Prediction, COVID-19, Regression, Linear, Multi Linear, Vaccine, Polynomial.

I. INTRODUCTION

The Coronavirus/COVID-19 pandemic of 2019/2020 is still taking its terrible toll as we write this [1] Coronavirus disease 2019 (COVID-19) may be a contagious infection caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). it had been first identified in December 2019 in Wuhan, China, and which has since spread globally, evolving into an ongoing pandemic. Common symptoms include cough, fever, fatigue, breathlessness and loss of smell and taste.

The World Health Organization (WHO) announced the outbreak a Public Health Emergency of International Concern in January and a pandemic in March [2]. The COVID-19 has largely impacted on all the sectors like economy, education, healthcare, logistics and mental health of people. The pandemic has caused severe global economic disruption and has led to the postponement or cancellation of many events. According to the World Trade Organization, the trade has been plunged due to the pandemic and is expected to fall between 13% and 32%. Many economic experts state that it might take 10 years to improve the economy to its normal state [3].

The WHO states that, COVID-19 has impacted significantly in the health sector for non-communicable diseases such as Cancer, Alzheimer's etc. Since there are no vaccines for this disease, it has become a humongous task and utmost priority for the healthcare department to prevent the wide spread of the disease [3].

Tests for the presence of antibodies *could* offer a way for people who can prove COVID-19 immunity to go back to work [10], [11]. There are, however, challenges concerning the biological premise of 'immunity': the strength and longevity of COVID-19 immunity after infection are matters of current debate and research, as are the sensitivity and robustness of the relevant tests [12], [13] and the race to develop a viable vaccine [14], [15].

With the help of predictive analysis and supervised learning, we can predict the future cases which might be helpful for taking much better preventive measures and precautions. The proposed model is shown in Fig. 1. Regression models characterize the relationship between variables by fitting a line to the noticed data. Linear regression models apply a straight line, while logistic and nonlinear regression models apply a curved line. Regression allows you to calculate how a dependent variable change as the independent variable(s) change. We have used 3 supervised machine learning models for the regression of the data. The data set after a chain of visualization seems to be linear and hence, we have used 3 fundamental regression models. Simple linear regression is used to estimate the relationship between two quantitative variables. We can use simple straightforward reversion when you want to know how firm the relationship is between two variables The worth of the dependent variable at a certain value of the independent variable

Multiple linear regression is used to calculate the relationship between more than one independent variable

and one dependent variable. It's used when we choose to know how firm the relationship is between two or more independent variables and one dependent variable. Polynomial Regression is a regression algorithm that represents the relationship between a dependent and independent variable as nth degree polynomial.

II. RELATED WORKS

After research and survey, we have found out a paper with similar kinds of work.

"Prediction of Covid-19 pandemic based on Regression" a journal written by professors [5] which uses regression. They have used date from, Centre for Systems Science and Engineering (**CSSE**) is a research collection centre housed within the Department of Civil and Systems Engineering (CaSE) of John Hopkins University. Which has released multiple forms of the dataset, and in this case, they have selected Time Series Dataset, which is updated every day. They have used the dataset collected from 01/22/20 to 06/22/20, which is available from their official GitHub website They have used Panda's package for data pre-processing. They have selected Support Vector Regression and Linear Regression. the evaluation of the model's performance is measured in the terms of R-Squared(R^2), mean squared error (MSE) and mean absolute error (MAE). The major disadvantage of SVR is that it cannot handle large data sets and hence its performance is less compared to LR Which even we are following in our project, but we have used polynomial regression, linear and multi linear regression.

"COVID-19 Future Forecasting Using Supervised Machine Learning Models" a journal written by professors [6], have used similar techniques and models, but with more research and experimentation. Even they have used the dataset used in the study has been obtained from the GitHub repository provided by the Center for Systems Science and Engineering, Johns Hopkins University. It contains daily time series summary tables, including the number of confirmed cases, deaths, and recoveries. All data are from the daily case report and the update frequency of data is one day The have selected models such as LASSO Regression, Support Vector Machine, Linear Regression and Exponential Smoothing. They evaluated by the performance of each of the learning models in terms of R-squared (R2) score, Adjusted R-Square (R2adjusted), mean square error (MSE), mean absolute error (MAE), and root mean square error (RMSE). They conclude that predictions on death rate and according to results ES performs better among all the models, LR and LASSO perform equally well and achieve almost the same R2 score. In comparison, SVM performs worst in this situation. They confirm death rate will be increased in future.

"Predicting the Probability of Covid-19 Recovered in South Asian Countries Based on Healthy Diet Pattern Using a Machine Learning Approach" a journal written by professors [7]. They have used Covid-19 Healthy Diet Dataset to perform the research work. The source of this dataset is kaggle.com. they have collected features from the collected dataset, and they are Country, Alcoholic Beverages, Animal Products, Cereals - Excluding Beer, Meat, Vegetal Products, Others, Recovered, Death. For exploratory data analysis Top 10 Covid-19 affecting countries are selected from the report of the website Worldometer coronavirus update April 2020. They selected Three Machine Learning Algorithms – Random Forest, KNN and SVM are used for the prediction of recovery rate. They concluded that we could battle with Covid-19 by adapting a healthy diet. A healthy eating style could help to combat the Corona Virus. they have made an analysis on the energy intake from different categories of food of different countries. They also visualized through bar charts the calories consumption from various types of food at top Covid-19 affected countries. According to the analysis they have observed that top Covid-19 affected countries on average consumes more animal products than South Asian countries while consuming more plant-based products (e.g., cereals, vegetables). By different bar charts we compared here the diet pattern of South Asian countries and top 10 Covid-19 affected countries. Through this research work we have come to a decision that when a patient consumes more plant-based products (e.g., cereals, vegetables) and less animal products then recovered percentage is more otherwise less. Their proposed methodology will help the people of South Asian countries to predict the recovery probability in early stage based on healthy dieting style

"Machine Learning based COVID-19 Cough Classification Models - A Comparative Analysis" a journal written by professors [8]. This paper demonstrates three machine learning classification models and determines the better classifier among these three models. The model has made use of 15 dominant features. The paper has employed a method of selecting features based on ranking different scores derived from the feature selecting algorithms. Dataset of 86 cough audios sampled at 44KHz out of which 54 were that of COVID-19 positive patients and 32 were that of healthy individuals was obtained from University of Cambridge. Due to this disparity in the number of positive samples and healthy samples, an addition of 46 healthy cough audio samples

from Coswara database from Indian Institute of Science and of 18 healthy cough audio samples taken from Free sound database was made, thereby increasing the healthy audio samples to 96 and the overall total to 150 cough audio samples. The audios files were recorded on both android as well as web-based systems through a 'Covid-19 Sounds app' released by the University of Cambridge and the cough audios comprise both forced and natural coughs. They have selected logistic regression, Random Forest classifier and SVM. The machine learning models are trained and tested by using k-fold cross validation (5-fold cross validation) and the average accuracy, recall, precision, and F1-score across all the folds are reported. They have used cough audios to detect if the person has covid or not. Accuracy measures the overall efficiency of the classifier considering both positive and negative classes equally. It describes what proportion of the total number of predictions were correctly classified. Hence, it does not distinguish between false-positive errors and false-negative errors and certain area of application like medical application would be more sensitive to false negatives than false positives. Therefore, Recall and Precision would be an effective metric of evaluation in this application. Based on the tabulated result, SVM performs better than Logistic Regression and Random Forest on all performance metrics, while Logistic Regression performs adequately. Random Forest yields high accuracy and precision although its learning curve indicates that the model performance can certainly improve with more training samples.

"Covid-19 Outbreak Modelling Using Regression Techniques" [9], they have used Susceptible-Infected-Recovered (SIR) Model which is based around the assumption that the transmission of the infectious disease through contact among three classes of adequately mixed populations. The feature selection methods which were used by them are correlation feature selection, mutual information feature selection and recursive feature elimination. The data was collected from https://ourworldindata.org/coronavirus-source-data for India for total cases each day for 190 days beginning from the onset of outbreak in India. They have also used Decision trees, linear and polynomial regression. However, it should be noted that the scales in the above two graphs are different. They have concluded that machine learning models were able to model the outbreak better than the standard epidemiological model. Regular regression models such as linear and polynomial regressions can be improved upon drastically in modelling predictions by addition of features and it is highly helpful to incorporate machine learning in the pandemic outbreak modelling because of the robustness of their algorithms and accuracy obtained in the prediction results.

"Prediction of COVID-19 Spreading Using Support Vector Regression and Susceptible Infectious Recovered Model" [16] they have used Support Vector Machine Regression (SVR) and Susceptible- Infectious- Recovered (SIR) method as comparison. Two scenarios were used in this paper, i.e., best-case, and worst-case scenarios of COVID-19 spread in Indonesia. Predicting the maximum daily case is difficult because of many factors affecting the spread. However, the two scenarios, i.e., best-case scenario and worst-case scenarios can be used to handle the uncertainty. Best-case scenario showed the end of pandemic begin at the end of 2020 (SVR) and beginning of 2021 (SIR-Model). SIR-Model showed the end of epidemic was not too different for best-case and worst-case scenario (January 2021). In the other hand, SVR model of worst-case scenario showed that the epidemic will be ended on 5 March 2021, longer than SIR model.

"Regression Analysis of COVID-19 using Machine Learning Algorithms" [17], he dataset available from the data repository for the "2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). They have selected Support Vector Regression and Polynomial Regression. Through our analysis can be concluded that the Polynomial Regression Algorithm as compared to the Support Vector Machine Algorithm, shows an accuracy of approximately 93% by predicting the rise in cases for the next 60 days i.e., for the months of July and August.

"Analysis and Prediction of COVID-19 in Xinjiang Based on Machine Learning" [18], In this experiment, the official data provided by the national health commission were used to select the data from July 16 solstice 28, and the later data were predicted. They used Polynomial regression. In this study, the transmission model of COVID-19 was modeled, and the regression curve was obtained by multinomial regression according to the daily number of coVID-19 diagnosed in Xinjiang. We can see that in the short-term prediction, the use of polynomial regression can better predict the number of new confirmed cases per day. At the same time, we can also see that the scale of the epidemic has been well controlled with the strong intervention of the state and people's enhanced awareness of self-protection.

"Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting" [19], they have used polynomial and linear regression Analysis of dataset is done using linear and polynomial regressions which

involved metrics like accuracy, R^2 score, and MAPE. They concluded that polynomial is better than linear regression. Forecasting is done using Tableau and the results are found to be satisfactory.

III. PROPOSED ALGORITHM

3.1 System Architecture

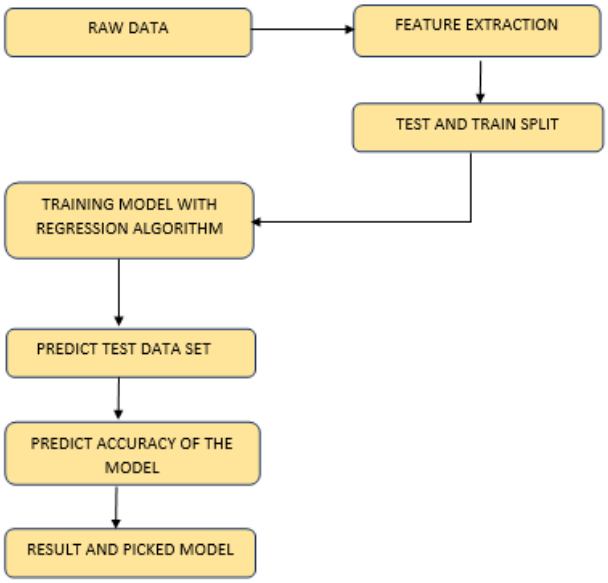

Figure 1: System Architecture

A. Raw data:

Raw dataset is taken from our world in Data website [4] which uses the most recent official numbers from governments and health ministries worldwide.

Feature Extraction:

Features are extracted by cleaning the data and fitting data for analysis by removing or altering data that's incorrect, fragmental, immaterial, duplicated, or inappropriately formatted. Since we are using regression models, we need an dependent and independent feature. So, the dependent feature will be number of people vaccinated on daily basis and independent feature will be considered accordingly based on the model. So, we only consider columns that are used and remove other columns and null values. We also remove null values

B. Train-Test split:

Train-Test Split Evaluation The train- test split is a strategy for estimating the performance of a machine learning algorithm. It can be used for regression or classification problems and can be used for any supervised learning algorithm. In this procedure, dataset is taken and divided into two subsets. The first subset is used to fit the model and is applied to as the training dataset. The successive subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This successive dataset is applied to as the test dataset. Train Dataset is used to fit the machine learning model. Test Dataset is used to evaluate the fit machine learning model. The main objective is to estimate the performance of the machine and choose the best fil model for the chosen dataset.

C. Use of Regression Algorithms:

Regression models characterize the relationship between variables by fitting a line to the noticed data. Linear regression models apply a straight line, while logistic and nonlinear regression models apply a curved line. Regression allows you to calculate how a dependent variable change as the independent variable(s) change. In this model, three different regression algorithms are used to predict the accuracy based on the test data set and the model which has the best accuracy and find the best fit model is considered. Here, we have used:

Simple linear regression:

It's used to estimate the relationship between two quantitative variables. We can use simple straightforward reversion when you want to know

Multiple linear regression

It's used to calculate the relationship between more than one independent variable and one dependent variable.

Polynomial Regression:

It's a regression algorithm that represents the relationship between a dependent and independent variable as nth degree polynomial. In this model, degree 2,3 and 4 has been used to see which degree best fit the model.

3.2 Flow Chart

A. Dataset Selection

Dataset is taken from our world in Data website [4] which uses the most recent official numbers from governments and health ministries worldwide. The population estimates which is used to calculate per-capita metrics are all based on the last revision of the United Nations World Population Prospects. It is updated each morning, with the most recent official numbers up to the previous day.

The data contains the following information:

The data (country vaccinations) contains the following information:

- Country- name of the country.
- Country ISO Code - ISO code for the country.
- Date - date for the data entry
- Total number of vaccinations - this is the absolute number of total immunizations in the country.
- Total number of people vaccinated – count of the number of people.
- Total number of people fully vaccinated - count of the number of people who are vaccinated twice.
- Daily vaccinations - for a certain data entry, the number of vaccinations for that date/country.
- Total vaccinations per hundred - ratio (in percent) between vaccination number and total population up to the date in the country.
- Total number of people vaccinated per hundred - ratio (in percent) between population immunized and total population up to the date in the country.
- Total number of people fully vaccinated per hundred - ratio (in percent) between population fully immunized and total population up to the date in the country.
- Number of vaccinations per day - number of daily vaccinations for that day and country.
- Daily vaccinations per million - ratio (in ppm) between vaccination number and total population for the current date in the country.
- Vaccines used in the country - total number of vaccines used in the country (up to date).
- Source name - source of the information (national authority, international organization, local organization etc.).
- Source website - website of the source of information

B. Data cleaning

Data cleaning is the process of fitting data for analysis by removing or altering data that's incorrect, fragmental, immaterial, duplicated, or inappropriately formatted. Since we are using regression models, we need a dependent and independent feature. So, the dependent feature will be number of people vaccinated on daily basis and independent feature will be considered accordingly based on the model. So, we only consider columns that are used and remove other columns and null values. We also remove null values.

C. Data visualization

Visualizing data using charts, graphs, and maps is one of the most impactful options to communicate complex data. Here we plot the graph for Date and Total vaccinations per hundred. We observed a linear correlation between Data and Total vaccinations per hundred. So, we decided to apply any of the regression algorithms. Here in this project, we have used matplotlib for plotting and visualizing the data.

D. Model selection

Supervised learning is the type of machine learning in which machines are trained using well labelled training data set, and on keystone of that data, machines predict the output. The labelled data indicates some input data is already tagged with the correct output. Regression algorithms are used if there is any relationship between the input variable and the output variable. It's used to predict continuous variables.

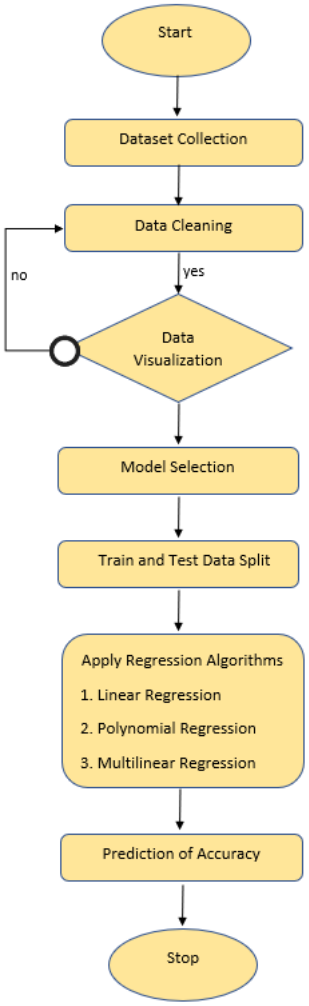

Figure 2: Flow Chart for the proposed model

Linear regression:

It's used to estimate the relationship between two quantitative variables. We can use simple straightforward reversion when you want to know.

It was selected because the graph for date and vaccination took by the people was linearly correlated. Since Linear regression works on linearly correlated data, this regression was chosen.

Polynomial Regression:
It's a regression algorithm that represents the relationship between a dependent and independent variable as nth degree polynomial. It was selected because the accuracy provided by linear regression was less. Polynomial regression is one of the types of linear regression in which the relationship between the independent variable x and dependent variable y is modelled as an *nth* degree polynomial.
Graphs in linear regression were in a straight line so it would not deliver a correct accuracy to a lot of the datasets hence polynomial regression was chosen where the graphs were on the curvier side. it establishes a curvilinear relationship.

Multi linear regression:
It's used to calculate the relationship between more than one independent variable and one dependent variable.
It's used when we choose to know how firm the relationship is between two or more independent variables and one dependent variable.
It was selected to test the prediction with several independent variables to predict the outcome of a dependent variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the independent variables and dependent variable. We have chosen date and country as independent variable to calculate the daily vaccination relation to the independent variables.

E. Train and test split

The train- test split is an approach for appraising the performance of a machine learning algorithm. In this procedure, the dataset is divided into two subsets. The first subset is used to fit the model and is bored to as the training dataset. The equivalent subset isn't used to train the model; instead, the input element of the dataset is furnished to the model, either prediction is made and compared to the expected values. This second dataset is referred to as the test dataset. We employed split percentages of train – 80% and test – 20%

F. Apply algorithm

Now we apply the algorithm one by one and check for accuracy. With linear regression, the accuracy was a bit low, compared to polynomial regression. Even the latter, was giving low accuracy with a degree 2 and 3 but accuracy was increased when the degree was 4. With Multilinear regression, we were able to include a lot of independent variables, so we were able to plot the graph in 3D.

G. Prediction of Accuracy

After the algorithm was applied now, we compare the performance of all the algorithm. We apply different regression algorithms to the dataset and measure the output by comparing the error rate of each algorithm and compare the accuracy. The lesser the error rate, the accuracy will be more and so then the best model that supports the dataset is chosen.

3.3 Algorithm

A. Simple linear regression

It's used to estimate the relationship between two quantitative variables. We can use simple straightforward reversion when you want to know how firm the relationship is between two variables. The worth of the dependent variable at a certain value of the independent variable

The formula for a simple linear regression is:

$y = \beta 0 + \beta 1 (x) + e$ ----------1

where:
$\beta 1$ = Slope of the line.
$\beta 0$ = y-intercept of the line.
x = Independent variable from dataset
y = Dependent variable from dataset
e = error term
the error in predicting the value of Y, given the value of X.

B. Multiple linear regression

It's used to calculate the relationship between more than one independent variable and one dependent variable. It's used when we choose to know how firm the relationship is between two or more independent variables and one dependent variable.

The formula for a simple linear regression is:

$y = \beta 0 + \beta 1 x1 + \ldots + \beta n xn + e$ ----------2

were
$\beta 1, \beta 2 \ldots \beta n$ = coefficient of input feature.
$\beta 0$ = y-intercept of the line.
x1, x2...xn = input feature
y = Dependent variable from dataset(output).
e = error term

C. Polynomial Regression:

It's a regression algorithm that represents the relationship between a dependent and independent variable as nth degree polynomial. The Polynomial Regression equation is given below:

$y= \beta0+\beta1x+ \beta2x^2+ \beta2x^3+.... + \beta nx^n$ ----------3

where:
β0, β1, βn = coefficient of input feature.
x, x^2...x^n = input feature
y = Dependent variable from dataset(output).

IV. RESULT AND IMPLEMENTATION

A. Dataset

Dataset is taken from our world in Data website, which uses the most recent official numbers from governments and health ministry's worldwide. The population estimates which is utilized to calculate per-capita standards are all rested on the last revise of the United Nations World Population Prospects. It's streamlined each morning, with the most recent official numbers up to the preceding day.

From the dataset all the dates and the number of cases is extracted and made into separate data frames. The dataset is then split into train and test data by dividing 30:70, 50:50, 60:40,80:20 percentage. For the actual prediction the dataset is split in the ratio of 80:20 train and test dataset. The dataset is split in such a manner to experiment and observe how the model learns with different quantities of training set using the evaluation parameters.

Machine learning models require all input and output variables to be numeric. This means that if the data contains categorical data, you must encrypt it to numbers before we can fit and evaluate a model. The Date column data is converted into a ordinal numbers and the column is considered as New_ID. For multi linear regression model, Country column data is converted into a ordinal numbers for further analysis.

B. Training

In this project, Jupyter Notebook and Google Collaboratory is used to execute Python code through the browser, as provides a free cloud service and supports free GPU. It is research oriented and does not require environment setup. It supports many machine learning libraries which can be loaded easily without any dependencies on hardware. The dataset is employed to validate the model.

The data is trained with train data and the test data that was taken from the dataset. The model is trained with parameters normalize and fit intercept set to TRUE and by applying different regression algorithms accordingly. Comparison of different regression algorithms that were used in this project are studied based on the accuracy for the effective prediction.

C. Evaluation Parameters

In this paper, the evaluation of the model's performance is measured by calculating the accuracy percentage for each model by use of R-Squared(R2) and mean absolute error (MAE).

MAE measures the prediction error. Mathematically, it is the average absolute difference between observed and predicted outcomes, MAE = mean (abs (observed - predicted))

R- squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model. R- squared explains to what extent the variance of one variable explains the variance of the second variable.

1. Simple Linear Regression:

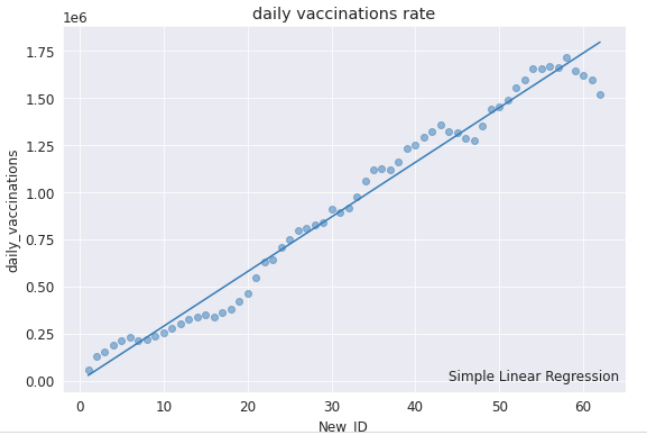

Figure 3: Daily vaccination measure using LR

The Date column data is converted into ordinal numbers and the column is considered as New_ID. Graph is plotted against New_ID as independent variable and daily_vaccination is predicted using the simple linear regression model. As observed the daily vaccinations are increasing on daily rate. The accuracy of this model is 97.759% as observed.

2. Polynomial Regression:

Degree = 2

The Date column data is converted into ordinal numbers and the column is considered as New_ID.Graph is plotted against New_ID as independent variable and daily_vaccination is predicted using the polynomial linear regression model. The accuracy of this model is 97.8533% for degree = 2 as observed

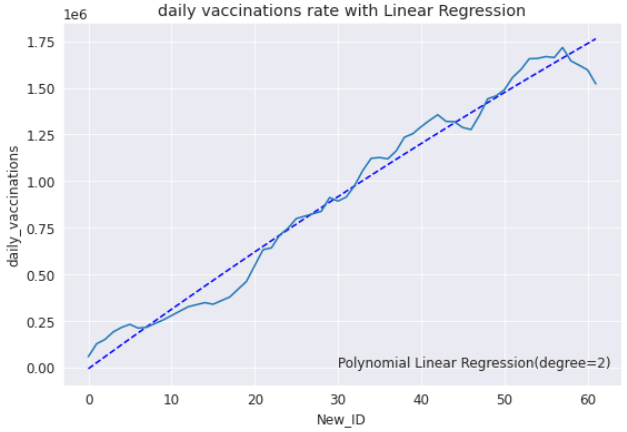

Figure 4: Daily vaccination measure using polynomial regression (degree= 2)

Degree = 3

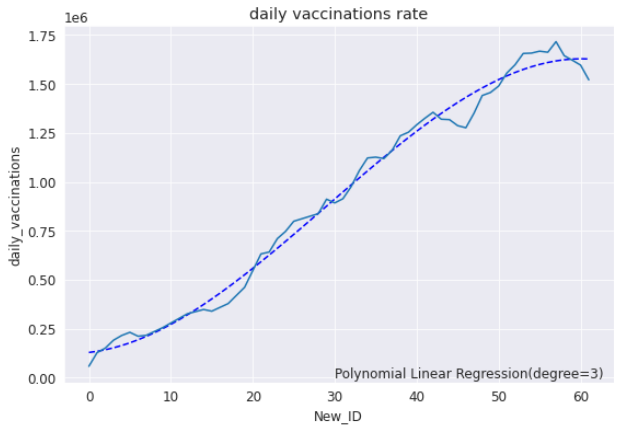

Figure 5: Daily vaccination measure using polynomial regression (degree= 3)

The accuracy of this model is 99.0016% for degree = 3 as observed.

Degree = 4

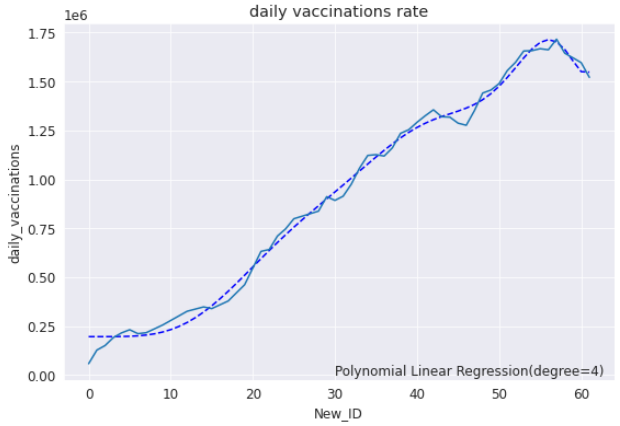

Figure 6: Daily vaccination measure using polynomial regression (degree= 4)

To fit the model with better accuracy, polynomial regression model with degree = 4 has been considered. The accuracy of this model is 99.0038% for degree = 4 as observed.

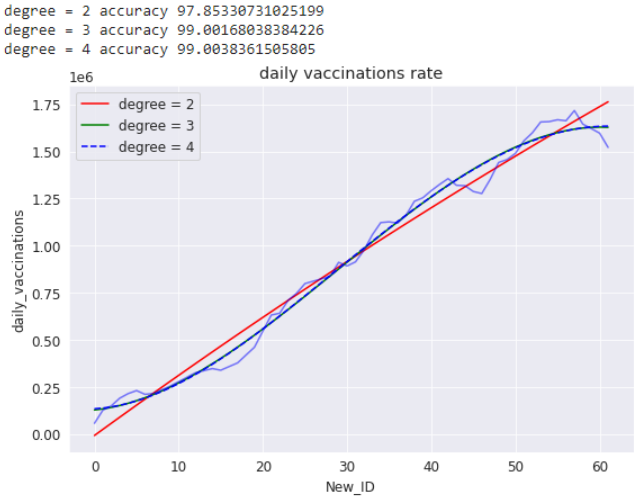

Figure 7: Daily vaccination measure using polynomial regression

As observed the daily vaccinations are increasing on daily rate.

The accuracy of this model is 97.8533% for degree = 2

The accuracy of this model is 99.0016% for degree = 3

The accuracy of this model is 99.0038% for degree = 4

To fit the model with better accuracy, polynomial regression model with degree = 4 is chosen.

3. Multiple Linear Regression:

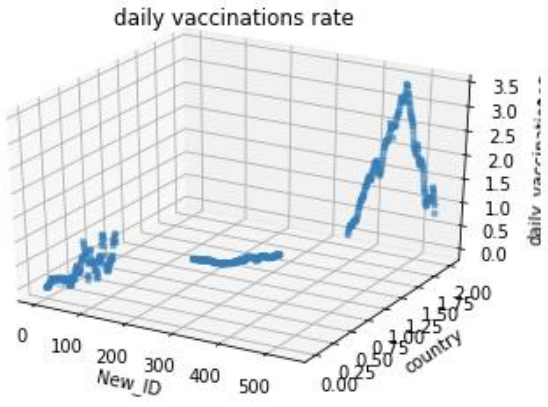

Figure 8: Daily vaccination measure using multi linear regression

Multiple linear regression is used to calculate the relationship between more than one independent variable and one dependent variable. Here we have used date and the country as two independent variable and daily vaccination as dependent variable. We have used considered Brazil, Canada, and United States countries for the graph above. The Date column data is converted into ordinal numbers and the column is considered as New_ID.

The Country column data is converted into ordinal numbers which is mapped as 0: Brazil, 1: Canada and 2: United States accordingly. Graph is plotted against New_ID, Country as independent variable and daily_vaccination is predicted using the Multiple linear regression model.

As observed the daily vaccinations are varying for different countries on daily rate. When most of the countries are considered, we can observe a positive linear corelation in the graph for the given data. The accuracy of this model is almost close to 99.98% with the error rate of 2.3295e-07 as observed.

When the performance of the model is evaluated against the R2 if the value is negative, it indicates that the model's performance is arbitrarily worse. And if the value is nearing to or is 1.0, the model is evaluated to be having best performance.

Here is the Evaluation Parameters for the model:

Model	R2	MAE	Accuracy(rounded)
Linear regression model	0.895009	20437.869	97.759
Polynomial Regression	0.999889	850.903	99.0038
Multi linear regression	1.0	2.3290327887411895e-07	99.98

Table 1: Evaluation Parameters for all the regression models

D. Comparison of Models

the model performs with over 90% accuracy. It is also observed that as the ratio of the training set is increased the model slowly improves its performance.

VI. CONCLUSION

In conclusion we can confirm that when compared the different regression models Polynomial Regression with degree 4 and Multilinear regression has better performance with good accuracy. Regression analysis is a dependable method of identifying which variables have impact on a subject of interest. The process of performing a regression allows us to confidently decide which factors matter most, which factors can be neglected, and how these factors influence each other. The current work can prove that the Covid-19 epidemic vaccination is growing linearly every day. This can be validated by visualizing Fig. 3. And Fig. 4. as the number of cases are rising in a straightaway fashion and proves that this will be a major risk until a stretch or two. Hence, we need to take the utmost preventives and measures to ease the spread of this sickness and get vaccinated. In future direction, deep learning models to predict covid19 vaccination to train the model with large data sets using TensorFlow.

REFERENCES

[1] "Coronavirus COVID-19 global cases by the center for systems science and engineering (CSSE) at Johns Hopkins University", [online] Available: https://gisanddata.maps.arcgis.com/apps/opsdashboard.

[2] Covid-19 Pandemic, https://en.wikipedia.org/wiki/COVID-19_pandemic Covid-19 Pandemic effects on world economy, https://www.wto.org/english/news_e/pres20_e/pr855_e.htm

[3] Covid-19 Pandemic effects on health sector, https://www.who.int/newsroom/detail/01-06-2020-covid-19-significantly-impacts-health-servicesfor-noncommunicable-diseases

[4] Dataset from official website of our world in data, https://ourworldindata.org/covid-vaccinations, github - https://github.com/owid/covid-19-data/tree/master/public/data

[5] Ashish U Mandayam, Siddesha S, Rakshith.A.C, S K Niranjan, "Prediction of Covid-19 pandemic based on Regression"

[6] "COVID-19 Future Forecasting Using Supervised Machine Learning Models" Available: https://ieeexplore.ieee.org/document/9099302

[7] Md. Showrov Hossen, Dip Karmoker,"Predicting the Probability of Covid-19 Recovered in South Asian Countries Based on Healthy Diet Pattern Using a Machine Learning Approach"

[8] Dr.Jayavrinda Vrindavanam , Dr. Raghunandan Srinath , Hari Haran Shankar , Gaurav Nagesh,"Machine Learning based COVID-19 Cough Classification Models - A Comparative Analysis"

[9] Ankita Bansal, Utkarsh Jayant, "covid-19 Outbreak Modelling Using Regression Techniques", Available: https://ieeexplore.ieee.org/document/9388347/

[10] "'Immunity passports' could speed up return to work after Covid-19", Available:

https://www.theguardian.com/world/2020/mar/30/immunity-passports-could-speed-up-return-to-work-after-covid-19

[11] "No 10 seeks to end coronavirus lockdown with ‘immunity passports’", [online] Available: https://www.theguardian.com/politics/2020/apr/02/no-10-seeks-to-end-covid-19-lockdown-with-immunity-passports.

[12] S. Malapaty, "Will antibody tests for the coronavirus really change everything?", Apr. 2020, [online] Available: https://www.nature.com/articles/d41586-020-01115-z.

[13] D. Male, J. Golding, and M. Bootman, "How does the human body fight a viral infection?", 2020, [online] Available: https://www.open.edu/openlearn/science-maths-technology/biology/how-does-the-human-body-fight-viral-infection.

[14] T. Thanh Le et al., "COVID-19 vaccine development landscape", Nat Rev Drug Discov, vol. 19, pp. 305-306, Mar. 2020, [online] Available: https://www.nature.com/articles/d41573-020-00073-5.

[15] N. Lurie, M. Saville, R. Hatchett, and J. Halton, "Developing Covid-19 vaccines at pandemic speed", N. Engl. J. Med., vol. 382, no. 21, pp. 1969-1973, 2020.

[16] Teddy Mantoro, Rahmadya Trias Handayanto, Media Anugerah Ayu, Jelita Asian,"Prediction of COVID-19 Spreading Using Support Vector Regression and Susceptible Infectious Recovered Model" Available: https://ieeexplore.ieee.org/document/9415858

[17] Ekta Gambhir, Ritika Jain,Alankrit Gupta,Uma Tomer,"Regression Analysis of COVID-19 using Machine Learning Algorithms" Available : https://ieeexplore.ieee.org/document/9215356

[18] Yunxiang Liu,Yan Xiao,"Analysis and Prediction of COVID-19 in Xinjiang Based on Machine Learning" Available: https://ieeexplore.ieee.org/document/9363798

[19] Saud Shaikh, Jaini Gala, Aishita Jain, Sunny Advani, Sagar Jaidhara, Mani Roja Edinburgh,"Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting" Available: https://ieeexplore.ieee.org/document/9377137