



**VIT**<sup>®</sup>  
**Vellore Institute of Technology**  
(Deemed to be University under section 3 of UGC Act, 1956)

**School of Computer Science and Engineering**

**J Component report**

**Programme : B.Tech**

**Course Title : Foundations of Data Analytics**

**Course Code : CSE3505**

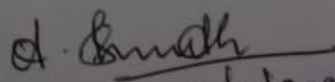
**Slot : F1**

**Title: Analysis of vaccination rate and its effect on spread of  
coronavirus**

**Team Members: Advait Deochakke | 20BCE1143**

**Aryan Shah | 20BCE1490**

**Faculty: Dr. Sheik Abdullah A**

**Sign:**   
**Date:** 18/11/2022.

## **Index Page:**

<b>Description</b>	<b>Page No.</b>
List of Figures	3
Abstract	4
Introduction	5
Literature review	6
Proposed methodology	8
Experimental results and discussion	9
Statistical analysis and interpretation	11
Visualization analysis	14
Conclusion	17
References	18

### **List of Figures:**

<b>Fig. No.</b>	<b>Description</b>
<b>1</b>	Proposed methodology
<b>2</b>	Correlation Matrix
<b>3</b>	Naive Bayes Analysis (Cases vs Tests, based on HDI)
<b>4</b>	Random Forest Variable Importance
<b>5</b>	Random Forest Predictions vs Testing values
<b>6</b>	Fraction of Deaths in the Population attributes to Covid
<b>7</b>	Vaccination Rate based on GDP (PPP)
<b>8</b>	Cases vs Vaccinations chart for Gaussian LM
<b>9</b>	Deaths vs Vaccinations chart for Gaussian LM

## **Abstract:**

The Covid-19 pandemic has resulted in dramatic loss of life around the world and created unexpected public health challenges. Vaccination is currently the only means of slowing the spread of the virus in the community. The main focus is the collection of data on vaccination available on the Internet. These datasets are cleaned and normalized for data analysis. The motivation is to provide necessary information about immunizations worldwide, as well as model older people.

Figuring out quantitatively whether the increasing rate of vaccinated individuals has had a definitive impact on the spread of novel coronavirus. If there is a notable relation between the two, to what extent they are interconnected how we can predict the rate of infection to change as time goes past and more people are vaccinated.

The venture ambitions to convey the evaluation of various ongoing vaccination programs around the world by using the use of the inferences determined from the scraped records from the internet. The python libraries used inside the exploratory facts analysis consist of NumPy, Pandas, Matplotlib, Seaborn, and Plotly.

### *Problem statement*

Whether there is a connection between vaccination rates and infection Rates?

What is the extent of said connection, if it exists?

Can we predict the decrease in infection rates as vaccination rates decrease?

Whether there is a notable relation between the two, to what extent they are interconnected, and how we can predict the rate of infection to change as time goes past and more people are vaccinated.

How does the spread of Coronavirus fare today?

## Introduction:

Covid-19 gripped the world in early 2020 and extensive research was conducted. Vaccinations were soon prepared, like the Moderna or Pfizer vaccine. Many people jumped at the chance to get these vaccines, and millions of people were vaccinated every day. Vaccine mandates, Mask mandates, and social distancing, combined with a prolonged period of social distancing and understanding from people has largely decreased infection rates.

As when the Covid-19 virus takes the arena in a shock. Covid-19 was declared as an endemic by using the world health organisation on 11th March 2020.

For stopping the spread of the virus diverse international locations, ought to put into effect a complete lockdown. This lockdown even though help in slowing the speed of infection but became a first rate element in other adversaries

like economic meltdown, activity loss, melancholy, and different low-budget and psychological problems. Scientists, academicians, and pharmaceutical research institutes labored tough toward developing a vaccine towards this virus. The Covid-19 vaccines

are supposed to offer immunity against the virus. The COVID19 vaccines are broadly credited for their position in decreasing the spread, severity, and death as a result of a coronavirus. Almost all countries have slowly shifted to a phased distribution plan for the vaccinations that prioritize the citizens at the highest danger of damages which could result from complications, together with the elderly, and those at a high chance of exposure and transmission, along

with healthcare employees. Seeing that most of the vaccines have been given emergency approvals, there have been various misconceptions associated with them. With the span of time, the acceptability of vaccines has elevated.

The nations which might be manufacturing these vaccines are exporting the same to the alternative international locations. in this project the dataset tracks the whole quantity of COVID-19 vaccinations administered in every usa, broken down by first and 2nd doses (in which country wide facts is available), and derived each day vaccination fees and population-adjusted figures.

Moreover, the pandemic has created a massive wealth of data due to the global-scale cooperation, creating opportunities for exercising various data analytics concepts.

## Literature review:

The paper discusses the case in its infancy for Bangladesh, where limited awareness and various socio-economic factors, combined with very high population density can lead to explosive growth different from global trends recognized and evaluated by population density, GDP, HDI etc into account. Advanced models such as LASSO, RNN, GRU, and VAE are taken into use. We have focused on implementing short-term forecasting models where accuracy is more likely to be achieved. The paper uses algorithms and models which have successfully predicted curves for numerous other.

The Prophet algorithm made incredibly consistent and good predictions across 35 different periods of rounds. Due to the nature of Linear Regression, there were obvious unconformities during the period of 9 march 2020 through April 2020. The prediction is almost spot on 3rd algorithm onwards with the PR algorithm and Poly-MLR giving the best results in round 3 and 4 of the iterations after SVR prediction continued to be slightly off from actual. Later from the 5<sup>th</sup> to the 34<sup>th</sup> Prophet algorithm gives the best results and confirms correctly to the real-world data.

A majority Covid-related papers have been published in the past year or two dealing with modelling the spread, including models involving segmentation and compartmentalizing of infected populace into sections of Susceptible, Infected, and Recovered (SIR Models). The model, after careful calculation of the various responsible factors, gives a terrifyingly accurate prediction of cases per day. High Caution and Low Sense of Safety, an ideal state to prevent greater spread of the disease.

Contrary to our previous discussed works, the current paper accounts for predicting whether an individual is infected or not based on symptoms experienced and various different categorization techniques, the authors showcase the Bayes theorem of probability and how it related to classification-based Machine Learning Algorithms. The 3 base algorithms of Decision Tree, Linear Regression, and Naive Bayes showed close to a 95% accuracy with ~85% sensitivity, ~90% specificity. Whereas, Artificial Neural Network and Support Vector Machine show closer to 90% accuracy, with ~92% sensitivity and ~80% specificity.

Rather than using extremely advanced concepts or having revolutionary results, it shows the very foundation of machine learning and how it can be applied to day-to-day tasks. Machine learning model is presented using

statistical visualization graphs to better predict the spread of the COVID-19 pandemic. Data used in this study were collected from official data repositories such as the official website of Johns Hopkins University, WHO and Worldometer. The core idea is to use supervised machine learning algorithms for time series forecasting. Formulas were developed by WHO to calculate confirmed rates of change and mortality, patient recovery rates, and pandemic growth rates, etc.

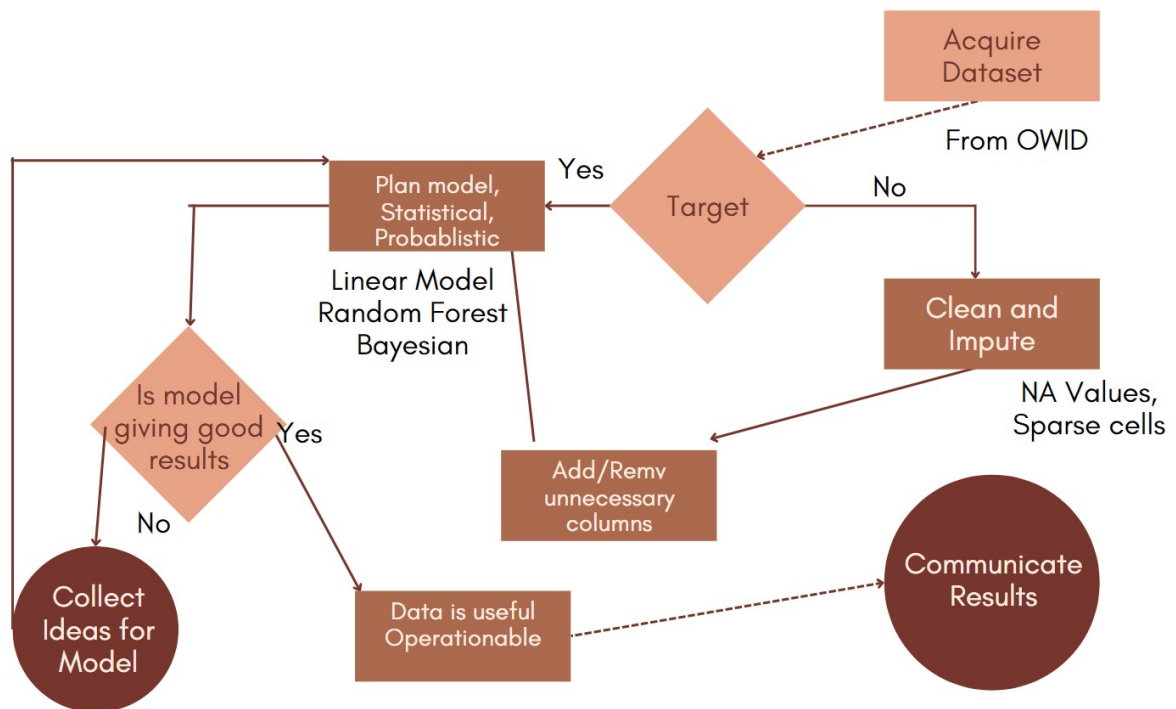
To validate the performance of the proposed method, they used the root mean square error for each of his three attributes: confirmed cases, recovery, and death. They showed the final predictions of the proposed model for all attributes. They conducted a comparative study using state-of-art methods. Their findings predicted a sharp decline in his COVID-19 infections worldwide in the first week of September 2021.

This paper presents an overview and brief analysis of the leading machine learning predictive models for COVID-19. The paper presented a keyword-based network visualization, and the top keywords researchers focused on, including coronavirus, forecast, epidemic, humans, statistical analysis, quarantine, hospitalization, mortality, and weather. Additionally, this paper will help researchers identify key gaps in the research field and develop new machine learning models for predicting COVID-19 cases.

Extrapolation of epidemic impacts from early stages of growth is affected by model, data, and behavioural uncertainties. Although the complexity and methodologies vary, simple models can still be used as a reference during the early growth stages of the epidemic and provide the basis for more complex models of infection to understand the course of disease development. In combination with monitoring results from serological studies, it helps modelling researchers to accurately calibrate epidemiological models based on real-world situations. Various studies have been conducted on the impact of contact tracing and social isolation, but it has been argued that improved quarantine and reporting rates, as well as the wearing of face masks, are essential for epidemic prevention and control.

## Proposed methodology:

Fig.1:





## **Experimental results and discussion :**

Approach lies along the lines of co-relating vaccine rates by identifying weekly averages of increasing and cumulative vaccines (segregated by markers such as country, month, wave, etc – TBD) to the infection rate in that same category, and analyzing the processed data.

A smoothed distribution line and colour coded bar graphs, among other visualization tools are used in the last few plot to highlight the trends in data and make them easier to see.

We utilize a Generalized Gaussian Linear Model to check how much of variation in data can be explained with a straight line, and it turns out that, much like one would have thought, not a lot.

Fitting a random forest regressor to the data set shows how much better models which account for complex relations can be over models which do not, as the random forest regressor shows well over 97% fit to the testing data, based on the training set used.

Meanwhile, we have also plotted the variation of vaccination and vaccination rates with time, cases, and deaths for countries in different income groups, and also with respect to GDP (PPP) and median age across countries.

We recognize that the data is imbalanced with respect to records, eg the number of deaths is orders of magnitude less than number of cases and tests, while features such as positive rate are numbers between 0 and 1. As such, we also scale the data to mitigate the differences due to such properties.

Furthermore, we use a NB Probabilistic model along with hierarchical clustering to generate tables for easy data interpretation. They easily show the variation in chances across sets of data at a glance, due to the intuitive approach utilized and easy-to-digest formatting.

**Full R Code:**

[https://github.com/Advait177013/FDA\\_JComp](https://github.com/Advait177013/FDA_JComp)

**Dataset:**

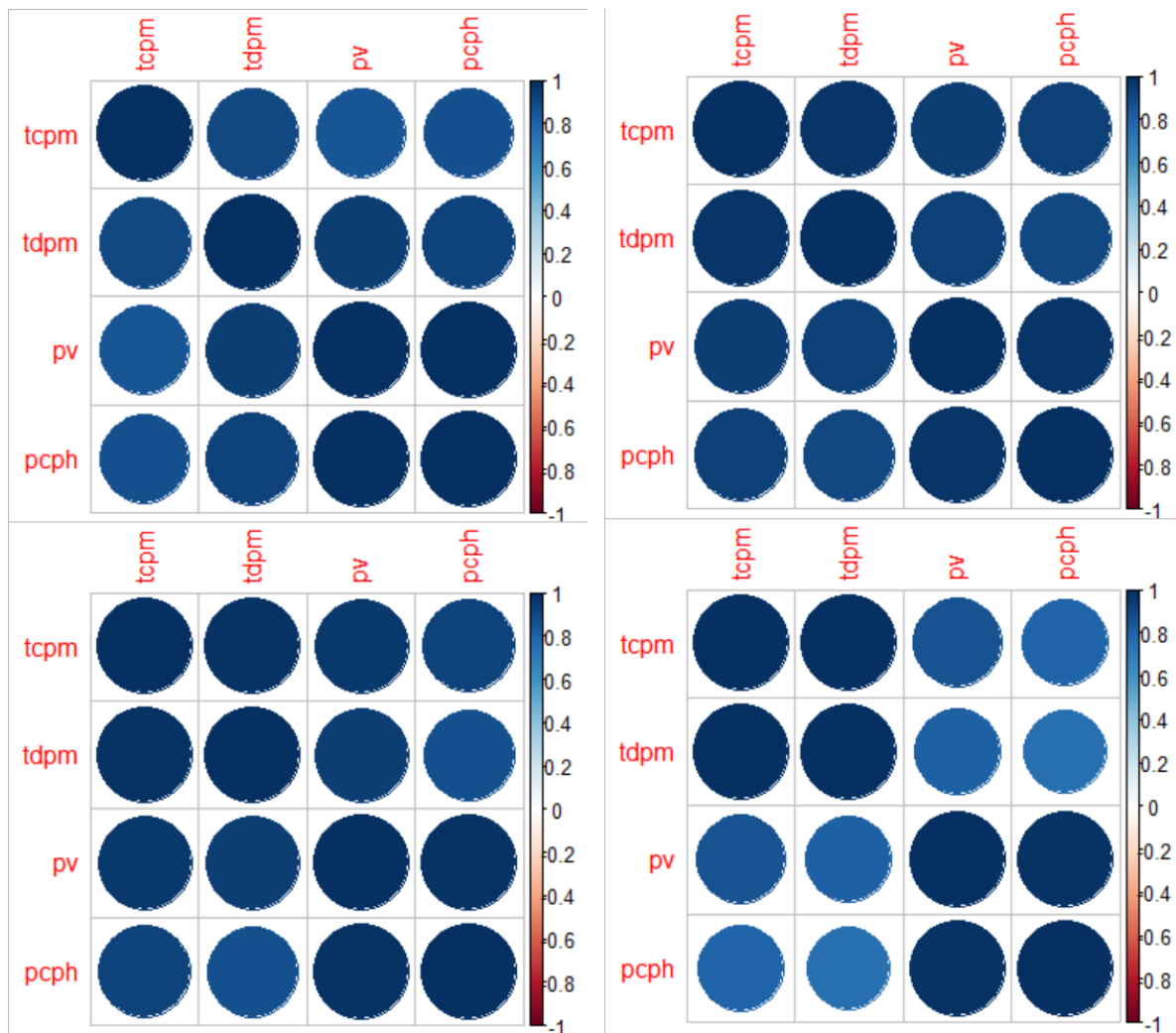
<https://ourworldindata.org/covid-vaccinations>

Dataset maintained by John Hopkins Uni. and Oxford University.

## Statistical analysis and interpretation

### Correlation Matrix for Income Groups

Fig.2:



As we can see, correlation between variables tends to be high, signifying that they can be used to explain each other, thus validating the input criterion for our model.

## Naive Bayes probabilistic model and querying

Fig.3

For high HDI countries, cases per million and tests per thousand

	cpm < 1k	cpm 1k-100k	cpm 100k-250k	cpm 250k-400k	cpm > 400k
<i>late 2021</i>	0.02	0	0.16	0.16	0.36
<i>early 2022</i>	0.59	0	0.43	0	0.25
<i>late 2022</i>	0.39	0.01	0.25	0.06	0.33

	tpt < 1k	tpt 1k-10k	tpt 10k-20k	tpt 20k-30k	tpt > 30k
<i>late 2021</i>	0.41	0	0.74	0.01	0.06
<i>early 2022</i>	0.57	0	0.04	0.15	0.02
<i>late 2022</i>	0.03	0.2	0.01	0.75	0.01

For low HDI countries, cases per million and tests per thousand

	cpm < 1k	cpm 1k-100k	cpm 100k-250k	cpm 250k-400k	cpm > 400k
<i>late 2021</i>	0.12	0	0.75	0	0.19
<i>early 2022</i>	0.86	0	0.15	0.08	0.02
<i>late 2022</i>	0.02	0.08	0.02	0.71	0

	tpt < 1k	tpt 1k-10k	tpt 10k-20k	tpt 20k-30k	tpt > 30k
<i>late 2021</i>	0.95	0	0.12	0	0
<i>early 2022</i>	0.05	0	0	0.85	0
<i>late 2022</i>	0	0.87	0	0.14	0

## Random Forest Prediction for Vaccinations based on available features

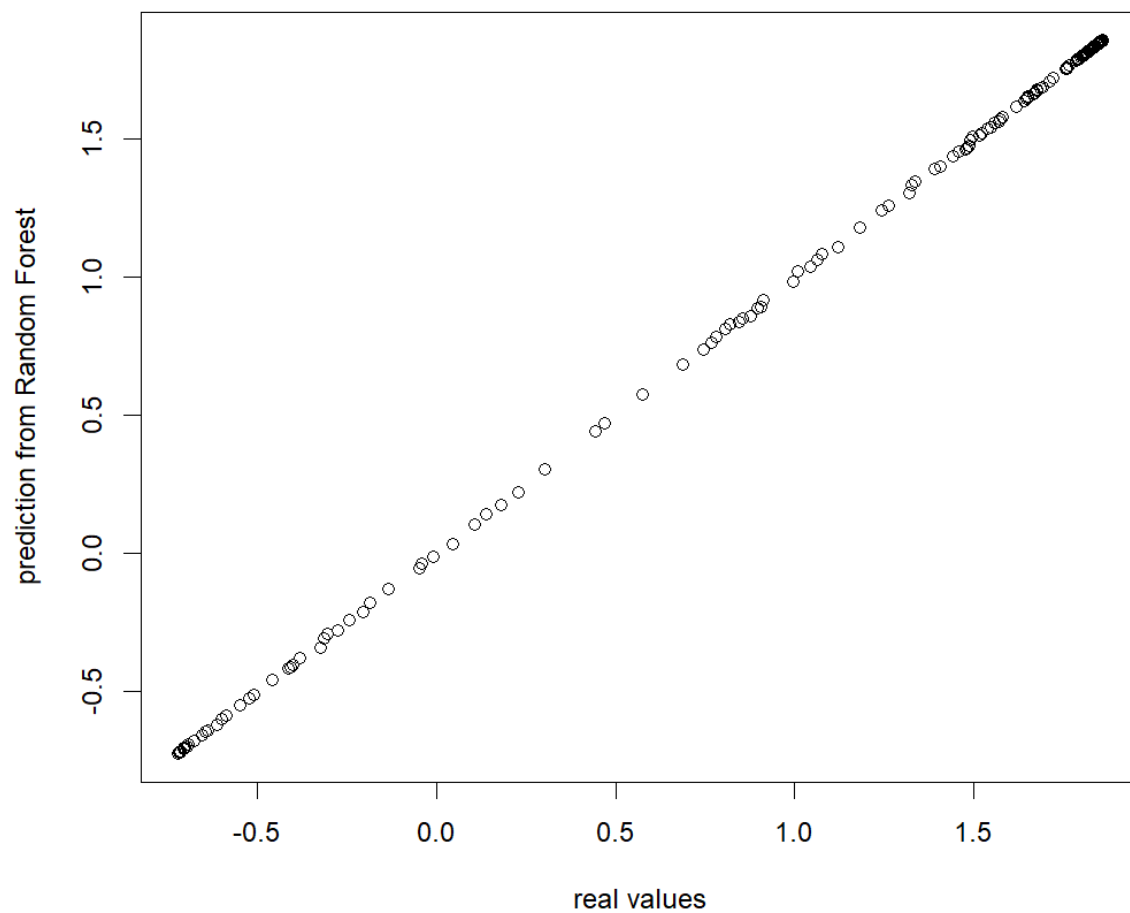
### Importance of Variables

Fig.4:

	%IncMSE	IncNodePurity
<i>total_cases_per_million</i>	18.34	310.34
<i>total_deaths_per_million</i>	18.3	316.48

Fig.5:

### Comparing Predictions to Real Values

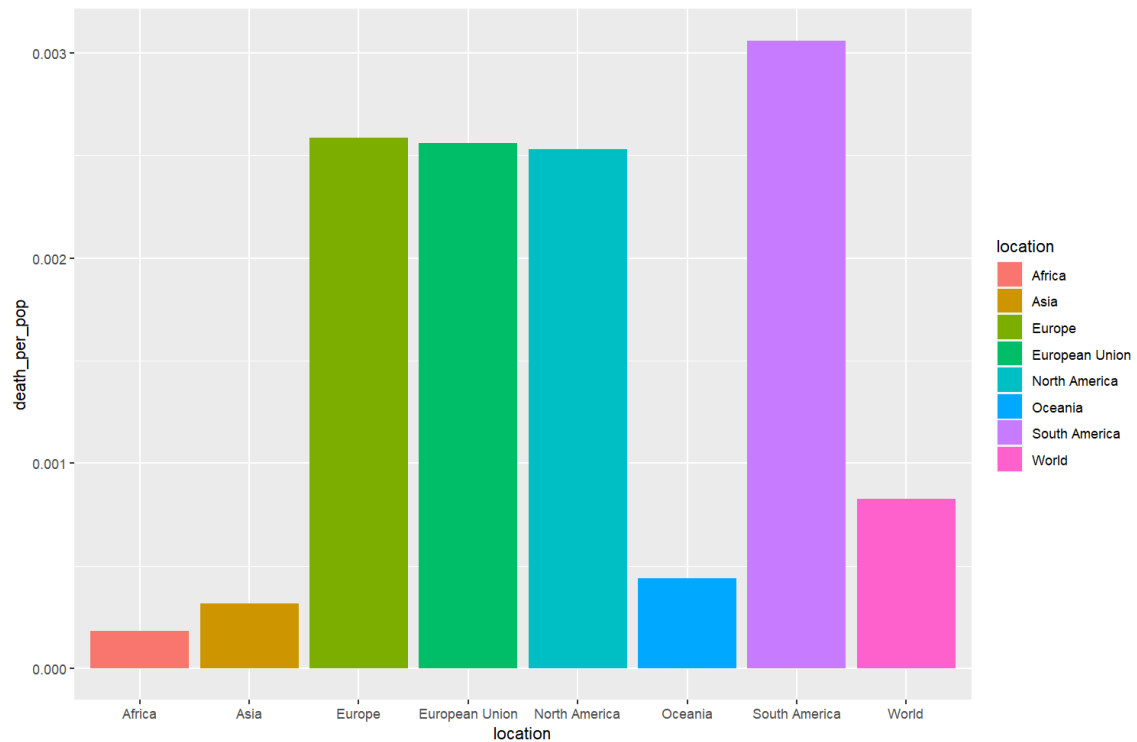


## Visualization analysis:

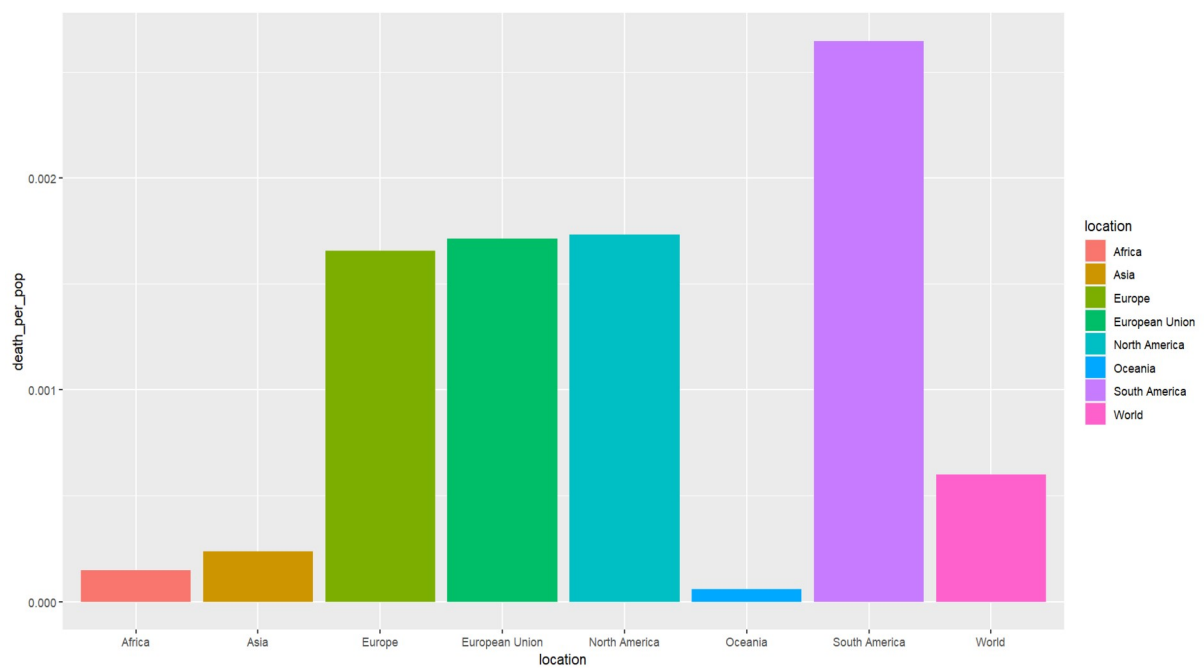
Fig.6:

### Fraction of deaths with respect to total population

September 2022



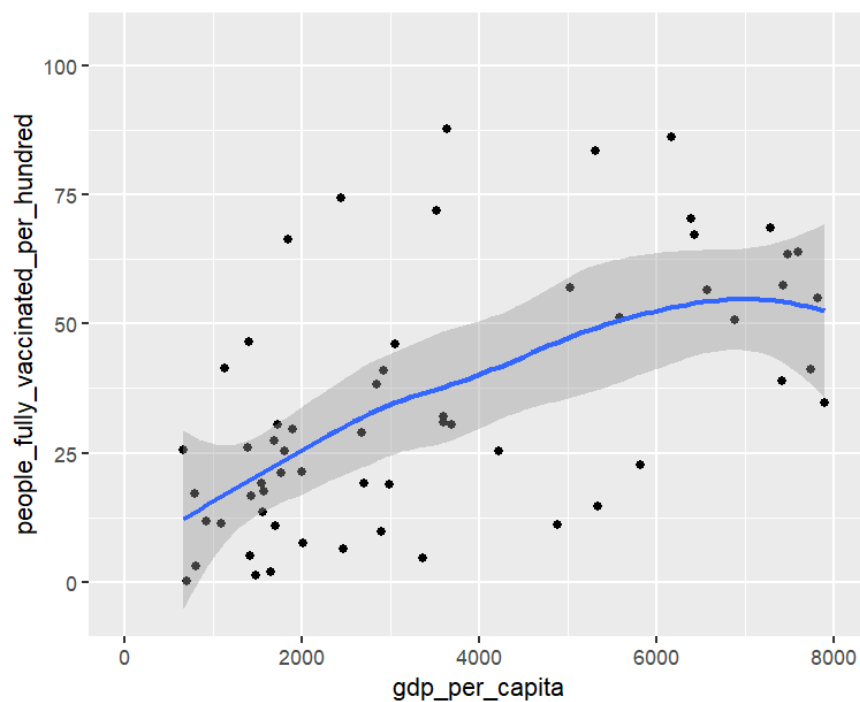
September 2021



## Vaccination rate based on GDP (PPP)

Fig.7:

September 2022



September 2021

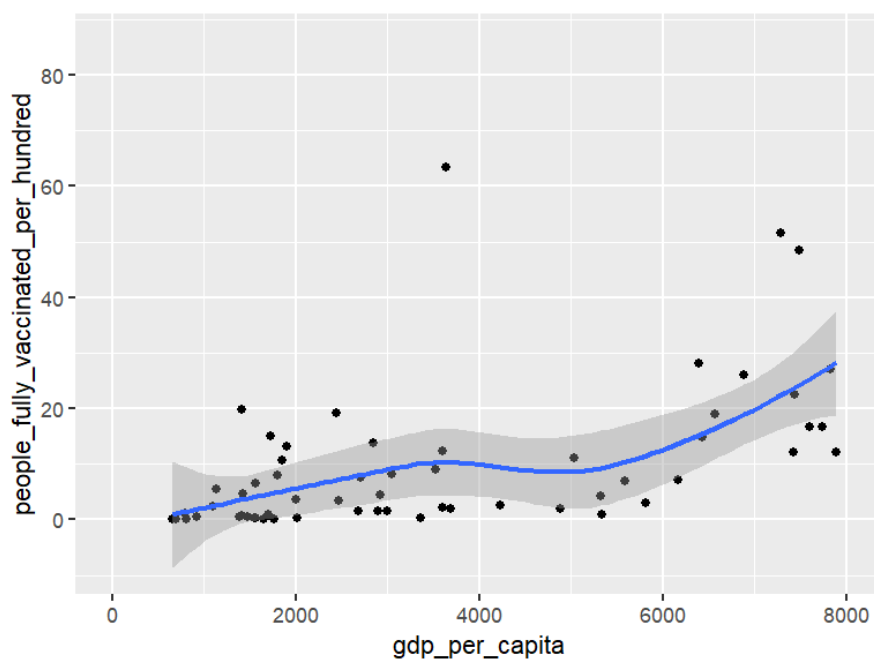


Fig.8:

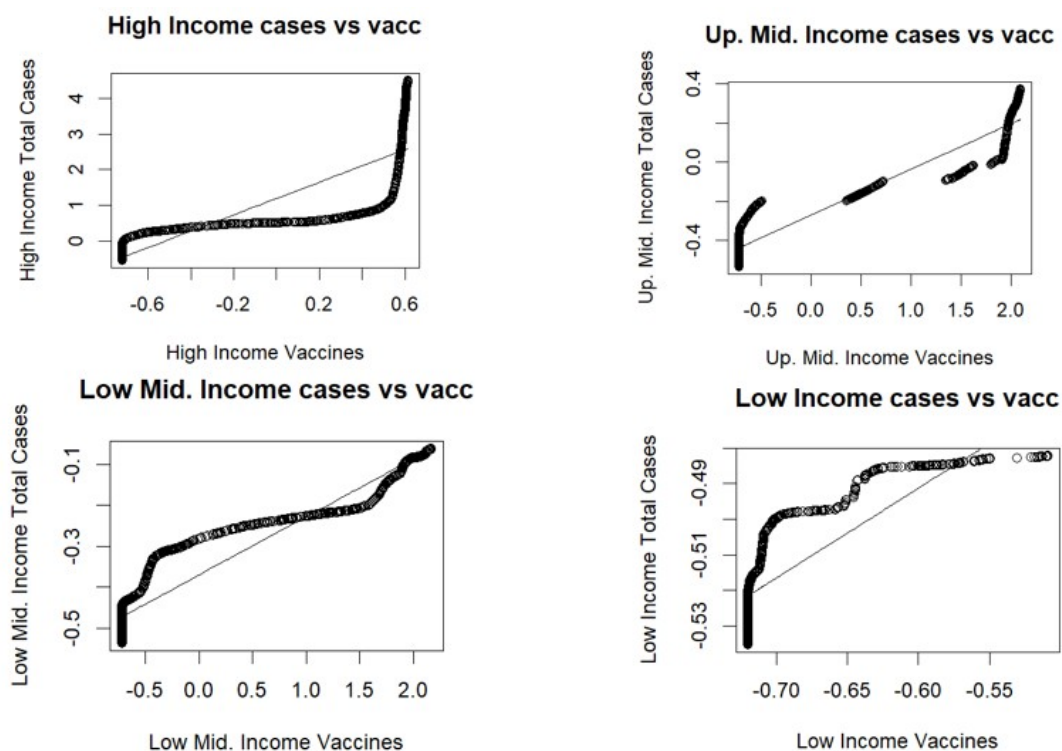
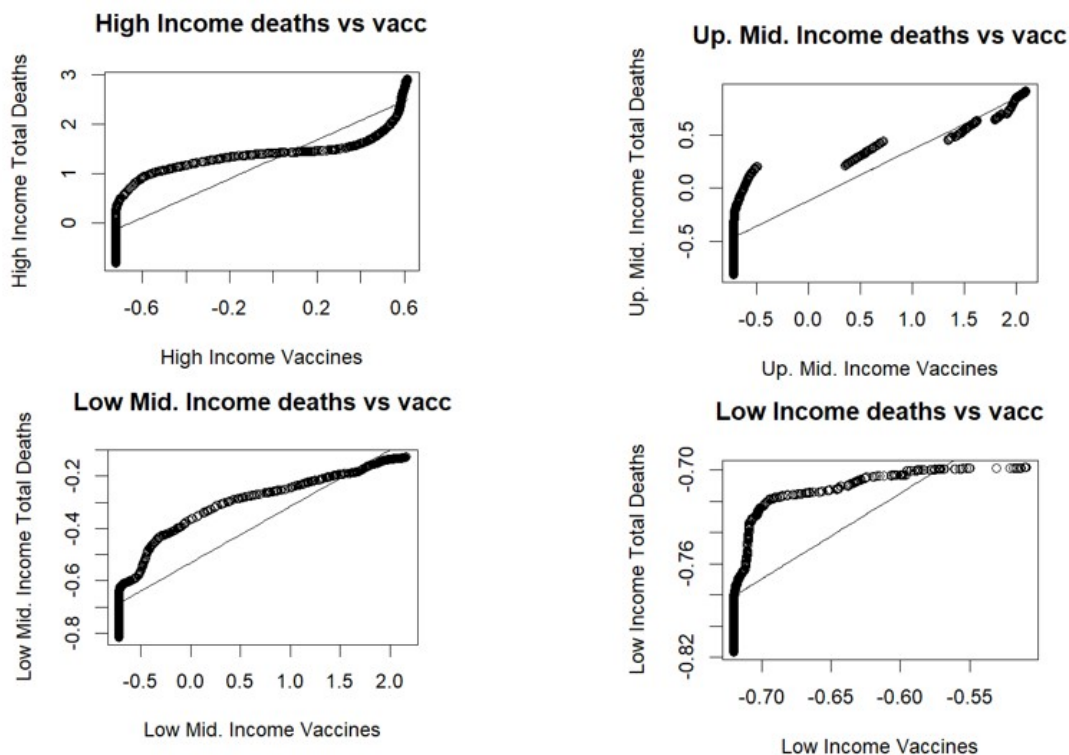
**Predicted vs Real values – Cases vs Vacc Gaussian Linear Model**

Fig.9:

**Predicted vs Real values – Deaths vs Vacc Gaussian Linear Model**



## **Conclusion:**

We may confirm the existence of a relation between vaccination rates in a particular demographic, and their relation of the spread of COVID-19 in the same demographic.

We may find to what degree this relationship is, and hope to investigate into whether this is causation type of relationship between the former and the latter, or a co-relation type.

The project was successfully developed and found to perform well. In addition, the data entered by the user and the name of the patient's disease are saved in a database, which can be used as a record of the past and used for future treatment, contributing to the simplification of health management. In this we can use machine learning algorithms to show that diseases can be easily predicted using different parameters and models. New features can be added to this project to make it more productive, reusable, flexible, and hybrid.

## References

- 1) Satu, M.S.; Howlader, K.C.; Mahmud, M.; Kaiser, M.S.; Shariful Islam, S.M.; Quinn, J.M.W.; Alyami, S.A.; Moni, M.A. Short-Term Prediction of COVID-19 Cases Using Machine Learning Models. *Appl. Sci.* 2021, 11, 4266. [https://doi.org/ 10.3390/app11094266](https://doi.org/10.3390/app11094266)
- 2) Usherwood, T., LaJoie, Z. & Srivastava, V. A model and predictions for COVID-19 considering population behavior and vaccination. *Sci Rep* 11, 12051 (2021). <https://doi.org/10.1038/s41598-021-91514-7>
- 3) Muhammad, L.J., Algehyne, E.A., Usman, S.S. et al. Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. *SN COMPUT. SCI.* 2, 11 (2021). <https://doi.org/10.1007/s42979-020-00394-7>
- 4) Malki, Zohair, et al. "The COVID-19 pandemic: prediction study based on machine learning models." *Environmental science and pollution research* 28.30 (2021): 40496-40506.
- 5) Rahimi, Iman, Fang Chen, and Amir H. Gandomi. "A review on COVID-19 forecasting models." *Neural Computing and Applications* (2021): 1-11.
- 6) Xiang, Yue, et al. "COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models." *Infectious Disease Modelling* 6 (2021): 324-342.
- 7) Chu, Xu, et al. "Data cleaning: Overview and emerging challenges." *Proceedings of the 2016 international conference on management of data*. 2016.
- 8) Austin, Peter C., et al. "Missing data in clinical research: a tutorial on multiple imputation." *Canadian Journal of Cardiology* 37.9 (2021): 1322-1331
- 9) [Villavicencio, C.N.; Macrohon, J.J.E.; Inbaraj, X.A.; Jeng, J.-H.; Hsieh, J.-G. COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA. *Algorithms* 2021, 14, 201. <https://doi.org/10.3390/a14070201>]
- 10) <https://onlinelibrary.wiley.com/doi/10.1002/jmv.27609> Correlation between vaccine coverage and the COVID-19 pandemic throughout the world: Based on real-world data Chao Huang, Lijun Yang, Jia Pan, Xiaomei Xu, Rong Peng- 'mass COVID-19 vaccination policy would protect the health and wellbeing of all, especially when the rate of vaccination passes 60%'
- 11) <https://doi.org/10.1016/j.msard.2022.104033> COVID-19 infection and vaccination against COVID-19 L. Pandit , A. Sudhir, C. Malli , A. D'Cunha - 'vaccinations are safe and significantly mitigates the risk of severe infection'
- 12) <https://doi.org/10.1016/j.meegid.2021.104834> Predictive analysis of COVID-19 eradication with vaccination in India, Brazil, and U.S.A Deepa Chaturvedi , U. Chakravarty 'we have predicted the possible timescales for the end of the epidemic for different values of vaccination rates'