Prediction of next wave of Coronavirus Disease 2019 (COVID 19) using regression model

Author

Saffiullah Khan

Reg. Number

206522

Supervisor

Dr. Urooj Fatima

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

September 2021

Prediction of next wave of Coronavirus Disease 2019 (COVID 19) using regression model

Author

Saffiullah Khan

Reg. Number

206522

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Software Engineering

Thesis Supervisor:

DR. UROOJ FATIMA

Thesis Supervisor's Signature:

____

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD
September 2021

Declaration

I certify that this research work titled "*Prediction of next wave of Coronavirus Disease 2019 (COVID 19) using regression model"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources has been properly acknowledged/referred.

Signature of Student
Saffiullah Khan
MS – 17 – CSE

Plagiarism Certificate (Turnitin Report)

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

Signature of Student
Saffiullah Khan
00000206522

Signature of Supervisor

Copyright Statement

- Copyright in the text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST College of Electrical and Mechanical Engineering (CEME). Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST College of Electrical and Mechanical Engineering (CEME), subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the CEME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST College of Electrical and Mechanical Engineering (CEME), Islamabad.

Acknowledgements

Firstly, The Most Gracious **Allah** has given me the strength and ability to write this report and complete our degree. We pray that Allah would keep on giving us strength to better this world and follow the footsteps of our role model Prophet Muhammad (p.b.u.h). Secondly, I would like to give my regards to **Dr. Urooj Fatima** whose unwavering support has been essential during this milestone. I am also very grateful for her continuous guidance and help.

I am profusely thankful to **Dr. Arslan Shaukat** and **Dr. Wasi Haider Butt** for an excellent guidance throughout this journey and for being part of my evaluation committee. I am thankful to my teachers who have guided me and molded my abilities so that I can succeed in my field. Lastly, I want to acknowledge my parents and siblings who lifted my spirits during the hard time. A very special thanks to all my friends, for always listening to me and being a sounding board.

Abstract

The Coronavirus Disease 2019 (COVID 19) has caused chaos everywhere in the world, and is prevailing with time. On 11th March 2020, World Health Organization (WHO) declared COVID-19 a global pandemic. Nonetheless, it is imperative to monitor the number of patients being affected. In this era of technological advancement and unprecedented change where the rate of innovations and discoveries are occurring at quite a pace, whilst in the field of health, a lot of work is still being done manually.

Many advancements have been made towards the domain of treatment nevertheless, no such initiative has been taken to automate the process of making predictions of deadly viruses until the year 2020 when the world was suffering through a pandemic as last pandemic, Influenza occurred back in 1918. Predictions with respect to waves of virus were initiated in 2020 when WHO made the covid-19 data available for researchers. Every country is facing issues while handling coronavirus in terms of lack of vaccinations, in adequate medical facilities, shortage of oxygen and logistics. This due to sudden outbreaks of coronavirus that does not give ample time to countries to make necessary arrangements as there is no mechanism to predict the upcoming waves. To overcome this problem, in this research we have aimed to generalize the approach towards the prediction of the next wave using multiple linear regression model to help international community take necessary actions to contain the spread at global level.

In our model total 17 attributes out of 113 were selected and 16 were independent variables (inputs) and 1 was a dependent variable (output). A study of trends of disease spread in the past and people affected by the disease daily are taken into consideration to predict the upcoming wave of COVID-19 globally. This will inform the region to investigate at country level to identify what Standard operating procedures are not being followed which will cause the predicted wave. Furthermore, prepare the countries of that region to arrange necessary medical equipment and take preventive measures to contain the spread in order to avoid the predicted wave. Moreover, this research fills many gaps highlighted during the previous researches such as handling of missing values in the input data given, skewed values, feature engineering and selection.

For this work, dataset is taken from Our World in Data (OWID). The number of active cases of a certain period along with 15 other attributes are given as the input to the model and linear regression model techniques are applied to the data in order to predict the upcoming

COVID-19 wave based on the analysis performed on new active cases and deceased cases with respect to dates.

Forecasting future trends will help international community to make necessary arrangement to contain the spread such as travel restrictions, focus of International community to expediating the process of vaccination and helping countries of a specific region which are unable to contain the spread. This pandemic is a global issue and apart from efforts being done by local governments of all countries, steps shall be taken on international platforms to eliminate this virus in a joint effort. The performance of the designed model is compared with the historic trends and other published methods, which demonstrates that the designed model provides predictions that are more accurate and precise up to 97.6%. The system is fast, efficient and has a high response rate as compared to the other models.

Key Words: *Coronavirus Disease; COVID 19; ML; Regression Model; Owid; EDA; Our World in Data; Multiple linear regression; R square; mse ;rmse.*

Table of Contents

Declaration.......1
Plagiarism Certificate (Turnitin Report).......2
Copyright Statement.......3
Acknowledgements.......4
Abstract.......5
Table of Contents.......7
List of Figures.......8
List of Tables.......9
CHAPTER 1: INTRODUCTION.......11
1.1 Background and Motivation.......12
1.2 Aims & Objectives.......13
1.3 Problem Statement.......13
1.4 Thesis overview.......14
CHAPTER 2: LITERATURE REVIEW.......17
2.1 Literature Review.......18
2.2 Critical Analysis.......25
CHAPTER 3: PROPOSED METHODOLOGY & IMPLEMENTATION.......28
3.1 Problem Definition.......28
3.2 Research Methodology.......30
CHAPTER 4: Results & discussion.......43
CHAPTER 5: Model Evaluation.......47
5.1 R square Evaluation:.......47
5.2 MSE Evaluation:.......48
5.3 RMSE Evaluation:.......48
CHAPTER 6: CONCLUSION & FUTURE WORK.......50
6.1 Conclusion.......50
6.2 Future Work.......50
REFERENCES.......52

List of Figures

Figure 1 “Flow chart of the methodology” 22

Figure 2 “Process of extracting and loading data” 23

Figure 3 “Process of reading data” 24

Figure 4 “Exploratory Data Analysis” 24

Figure 5 “Process of cleaning numerical data” 25

Figure 6 “Process of cleaning categorical data” 25

Figure 7 “Transformation applied on numerical features” 26

Figure 8 “Feature Scaling applied on categorical features” 27

Figure 9 “Feature Selection on the data” 28

Figure 10 “Regression Model” 29

Figure 11 “Results of the regression model” 31

List of Tables

Table 1 : Critical Analysis **18**

Table 2 : Results of regression model **31**

Table 3 : Comparison of models **32**

Chapter 1

Introduction

CHAPTER 1: INTRODUCTION

Pandemics have occurred throughout time, taking millions of lives worldwide. The first case of the novel coronavirus whose origins trace back to Wuhan city, China was first reported in 2019 [1]. The Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) has taken millions of lives all over the world; spreading through respiratory droplets when a person carrying the infectious disease coughs or sneezes close to others [2]. Moreover, ten million people are infected with over 500 thousand deaths within the six months since the origin of the virus, the rapid transmission of the virus has led the World Health Organization (WHO) to label it as a Global Pandemic [3]. The wide prevalence of the virus has affected lives all over the world, in addition to the development of vaccines, many challenges are being faced such as; availability and affordability of vaccines across the world, reliability of vaccines [4].

Coronavirus disease 2019 (COVID-19) has altogether affected the whole present reality and slowed down normal human exercises in a particularly remarkable manner that will have an extraordinary impression on the historical backdrop of humankind. Various nations have received various measures to assemble flexibility against this hazardous disease. Be that as it may, the exceptionally infectious nature of this pandemic has tested customary medical services and treatment rehearses. Along these lines, artificial intelligence (AI) and machine learning (ML) open up new components for powerful healthcare during this pandemic. AI and ML can be valuable for medication improvement, planning proficient determination techniques, and delivering forecasts of the disease spread to make necessary arrangements for logistics and plan smart lockdowns to contain the spread. These applications are profoundly subject to real-time observing of the patients and viable coordination of the data, where the Internet of Things (IoT) assumes a key part. IoT can likewise assist with applications, for example, automated drug delivery, responding to patient queries, and tracking the causes for disease spread. These days, we are noticing fast progression in each part of science and innovation. With the assistance of machine learning strategies, accessible information can foresee complex future patterns. With legitimate and adequate training information, a machine learning contraption can be used to concentrate on patterns and foresee the future patterns [5]. The proposed project is a machine learning approach which will be used to predict the next wave of Coronavirus Disease using a regression model. The system will predict the future trends of this deadly disease and will help in taking preventative measures at administrative level to make necessary arrangements to stop the spread of the virus and proactively plan the national level activities accordingly. This system, I believe will be an advancement towards more accuracy will lead to more accurate predictions. It will help the targeted audience to design better treatment plans for the affected patients and to take preventive measures.

1.1 Background and Motivation

Developments in technology are causing tremendous changes in many life sectors, may it be education, work, industry, or transportation. In the domain of health, computerization is still a work in progress. A lot of work has been done to develop such tools that may help with dealing with various diseases, for example, virtual reality simulations are used to help people overcome their phobias. With digitalization upsetting each industry, including medical services, the capacity to catch, share and convey information is turning into a high need. ML, big data and computerized reasoning (AI) can assist with tending to the difficulties that tremendous measures of information present. ML and AI can likewise help medical services associations fulfill developing clinical needs, further develop activities and lower costs. Moreover, Machine learning advancements can help medical services professionals distinguish and treat sickness all the more proficiently and with more accuracy and customized care.[6]

With humongous amounts of data being generated there is a need to effectively process it by saving a strenuous amount of time and work. Data is available on the internet in various formats and is easily accessible; however, to manipulate it and to derive a logical explanation from it takes a lot of human effort. To minimize this human effort in order to study the trends of the past and the future a machine learning approach can be applied. In machine learning different algorithms can be applied to reach an optimal solution. The process of manipulating the data is very transparent. Large amounts of data with millions of data points can be easily visualized and interpreted. Machine learning and IOT can be used to mitigate the covid-19. ML generated models can help in battling any upcoming pandemic. We can reuse the existing models with slight variation and train them on new types of datasets to mitigate the spread of any future pandemic. Moreover, the performance of these ML models can be improved by incorporating new machine learning methods. The performance gaps can be filled by revising the models to incorporate new advancements in machine learning.

1.1.1 CoronaVirus

The coronavirus belongs to a family of viruses that can cause illnesses in the respiratory system of the human body. Common symptoms for coronavirus are fever or chills, cough, shortness of breath or difficulty breathing, tiredness, muscle or body aches, headaches, loss of taste or smell, sore throat, congestion or runny nose, nausea or vomiting and diarrhea [7]. There are many types of Human coronaviruses amongst which the most common types are 229E (alpha coronavirus), NL63 (alpha coronavirus), OC43 (beta coronavirus) and HKU1 (beta coronavirus). Other human coronaviruses are MERS-CoV (the beta coronavirus that causes Middle East Respiratory Syndrome), SARS-CoV (the beta coronavirus that causes severe acute respiratory syndrome) and SARS-CoV-2 (the novel coronavirus that causes coronavirus disease 2019)[8].

Coronavirus is diagnosable through a laboratory test in which the healthcare provider collects a sample of your saliva or swabs your nose or throat. Coronavirus is treatable with Remdesivir (Veklury®) which is a FDA approved drug to treat patients infected with the COVID 19 infection [7]. Till date an approximately 215 million cases have been reported and 4.6 million have died by contracting this virus.

1.2 Aims & Objectives

Major objectives of this research are as following:

- Extract data from covid-19 dataset
- Perform Exploratory data analysis
- Cleaning and munging of data
- Feature Engineering on numerical data
- Feature Engineering on categorical data
- Feature Selection
- Regression Model Implementation
- Prediction of next wave based on countries

1.3 Problem Statement

Ever since the advancement of technology, a lot of expectations reside in the automation of artefact may it be in the field of education, aviation, psychology, teaching or health. During the former times, due to the lack of development and awareness our ancestors have suffered through droughts, famine and pandemics. Since then, a lot of work has been done specifically to help mitigate the severe effects of all kinds of real world challenges. Many advancements have been made specifically in terms of treatments however, no ground breaking work has been done to automate the process of predicting future occurrences. There was no sure way to know about the eventuality of an oncoming disaster until the emergence of machine learning. In the course of 2020, when the world was suffering through a pandemic from covid-19, a large number of lives were influenced straightforwardly from the infection. People are increasingly turned to clinicians when they develop or get ill, 4th waves are ongoing in several areas of the world, and the capacity limitations identified were not fixed completely despite many fatalities during initial waves of Covid-19. A number of regions face these challenges but nonetheless boast ongoing second waves. Although a lot of work has been done in order to control the spread of this deadly virus, there is no sure way of knowing if the world is going to suffer another wave.

Moreover, poor and labor class that are already living through hard times are affected by the lockdowns imposed the governments. The internment is severely touching the poor and migrant labours. Staying indoors might not be a viable alternative in current times while taking in account the risk that plenty of humans can also additionally die out of starvation and different ailments. International media is reporting regarding the pandemic and the way it's far effecting the lives of human beings. Many studies are being conducted in any respect ranges to fast collect information, come up with mitigation equipment and strategies and also implementation of the same. Therefore law makers and institutions need to have a holistic view of the latest situation and need to visualise the quantity at which the coronavirus can spread and cause deaths to enable them to make robust policies and plan next desirable steps that shall be taken to contain the spread.

For this purpose, there is a dire need for a machine learning model that predicts the future of the occurrence of the next wave. Predicting the upcoming wave for a specific country or a region with help of a model would help mitigate the risk of uncontrollable spread and proactively make the necessary arrangements to contain the spread.

1.4 Thesis overview

Chapter 1: Includes an overview of the thesis and introduction to background knowledge, domain information, motivation for topic selection, goals of research and structure of the thesis. Chapter 2: includes the detailed systematic literature review. Chapter 3 highlights the proposed methodology and implementation of our proposed idea. Chapter 4: includes a discussion of the achieved results and validation of these results from domain experts. Chapter 5: includes a conclusion section that concludes the thesis. Moreover, there are ideas listed for further improving the existing work under the heading of future work.

Chapter 2

Literature Review

CHAPTER 2: LITERATURE REVIEW

During the course of time numerous techniques have been applied for classification and predictions in various fields such as medical domain, engineering, and education. In this era of technological advancement and unprecedented change where the rate of innovations and discoveries are occurring at quite a pace, whilst in the field of health, a lot of work is still being done manually. Many advancements have been made towards the domain of treatment nevertheless, no such initiative has been taken to automate the process of making predictions of deadly viruses until the year 2020 when the world was suffering through a pandemic. Moreover, numerous machine learning models have been built to predict the outcome of people suffering from covid or the technologies that can be used to help mitigate the coronavirus. In this chapter overviews of several machine learning models have been presented and discussed.

2.1 Predictive modeling

Predictive modeling is a technique in which data mining and machine learning are used to predict future trend or patterns with help of historical data or existing data. Predictive modeling is a statistical technique and it works by analyzing the provided input and the recent data and projecting what it learns on a model while training to predict the likely instances for the future. Predictions by analyzing the historical data is helping every domain whether it is relevant to credit risks, customer purchases, medical, corporate sales etc.

Predictive models are not fixed and they are revised frequently to train those models on latest data and incorporate changes in the data. Assumptions are generated based on the past activities and current situation. Most predictive models are efficient and perform calculations in real time to help out companies, bank, administrative institutions and medical institutions in making decisions with respect to predictions generated based on past data. There are multiple types of prediction models and predictive analysis tools use different models with respect to nature of data and desired outcomes. Most commonly used types are (i) classification model, (ii) clustering model, (iii) Forecast Model, (iv) Outliers model and (v) Time series model. In this study we are using regression models.

Regressing models are one of the forecasting models which take historic data as input and provide desired outcome predicting future possibilities. Regression analysis is a predictive model technique which determines the relation between Predictor variable (Independent) and target variable (dependent). A curve line or a straight line is fitted to the data points in such a way that the distance of each data point from the line or curve is minimum. The benefit of using regression analysis is that is indicates both the significant relationships among the dependent and independent variable, moreover, it also shows the strength of impact of multiple independent variable on dependent variable.

Types of Regression:

(i) Linear Regression

(ii) Logistic Regression

(iii)Polynomial Regression

(iv)Ridge Regression

(v) Lasso Regression

To select right type of regression model we have to explore data and identify relation and impact of variable. Multiple metrics for model evaluation such as R-square, Mean square error, Root mean squared error can be used to determine the goodness of fit. Moreover, cross validation can be performed to evaluate your model to see if it fits and generate optimal predictions. Linear regression model is one of the important and readily used model. If the linear relation among the dependent continuous variable and multiple independent variable exists , such model is known as Multiple Linear Regression.

2.2 Literature Review

The prevailing situation of a pandemic has affected the conventional way of living. Moreover, numerous researchers have proposed methods to predict and forecast the next spike of the virus, take preventive measures, and handle the situation more effectively.

In the paper [9], the authors have proposed a method to forecast the number of totals, cumulative number of confirmed cases, number of recovered cases, and number of daily cases. The dataset used for this study was taken from Kaggle, which was based on 61 days. Furthermore, it was preprocessed using a Standard scaler for scaling the data, and then Support Vector Regression was applied for the prediction. In information pre-processing section, author has created the variable column (y) that is defined as dependent and independent variable (X) is set to be No. of days from start of March (i.e 1st March). X as independent variable is a numpy array of elements one to sixty one. 60% of the data is then taken for training and 40% of the data is considered for testing of the model. Segregation of data is done using python library function. After splitting the data set Standardscaler() function is used by the author for

standardization of both X and Y for testing and training. New objects are then created for standardized data of X and Y. To fit the object into the data for transformation of X and Y to achieve the optimal state ranging from -3 to +3 fit_transform() function is used. The data scaling is achieved, and regression application is done afterwards for prediction of new active cases. SVR is then applied as it fits for both linear and not linear. SVR regression model is used from Support Vector Machine class of sklearn python library. The performance metrics included: mean square error, root mean square error, regression error, and percentage accuracy. The suggested model had an accuracy of 99% for predicting deaths, cumulative number of confirmed cases, recovered cases, and the accuracy for predicting daily new cases was 87%.
The anticipated method forecasts the number of coronavirus active cases, number of everyday new cases, total demises along with number of day-to-day new deceases. Forecasting for patients that have recovered is also performed. With help of SVR, a machine learning model and Considering the past patterns, the upcoming pattern are projected. The Support Vector Regression is said to have performed better with a consistency in forecasting in comparison to other regression models such as logistic, linear and polynomial. The inconsistencies present in the dataset are catered using the methodolgy proposed. Measures for containment such as social distancing and hygiene shall be taken to stop the spread of coronavirus disease which is significantly higher as of now and needs to be controlled to lower the rate of progression by reducing the spikes reflected in the dataset. The authors have claimed the model to be a better prediction model than regression. However, the model is only based on the situation in India and has more accuracy in predicting deaths by taking into account number of active cases.

In order to speculate the trend of cases of coronavirus in the next upcoming days, the authors' Rath, S. et al have suggested a regression model [10]. A study regarding day-to-day numbers for people being affected by coronavirus disease is considered in order to forecast and predict future patterns and trends with respect to the active cases in India. The data set used is taken from WHO’s site from March 2020 to July 4, 2020. Correlation between the active cases, number of deaths and number of confirmed cases is determined and stated in the study conducted. The authors applied multiple regression models such as linear and multiple linear regression techniques to identify the patterns and trends of the affected number of cases using the dataset acquired from WHO.   Firstly, the data was cleaned for two data sets i.e. handling missing values. After data cleansing an analysis for correlation using Spyder of Anaconda Navigator application is performed in Python programing. Next, to determine the estimate for relative impact of new coronavirus cases due to active cases, linear regression models were applied on data pertaining to India. Considering the number of positive + cases (X) and No. of day-to-day active cases as (Y), the objective was to fit a regression line that would predict Y for X using linear regression model. The author highlighted the limitations that linear regression has as it sometimes discovers a relation between mean of values that are given for dependent and independent variables. As mean value cannot be considered as a complete picture of individual variable so linear regression does not provide a clear understanding of relationships between all variables. To address this issue analysis for more then one variable is done by using multiple linear regression model. In MLR model the dependent variable is the target variable and output of which is dependent on more then one independent variables. Multiple linear regression is then applied to the same data set with same segregation of data set as 80% for training the model and 20% testing purposes. Finally, The model evaluation matrix results are presented by the author which shows a comparison of both simple linear regression and multiple linear regression models. R2 score of the multiple linear regression and linear regression was 0.99 and 0.79 for data set pertaining to India. whereas, R2 score of multiple linear regression and simple linear regression was 0.99 pertaining to data set for Odisha. Thus, the commendable performance of the proposed method 2 (multiple linear regression) implies a strong prediction, however, the outcome is only for India and Odisha.

Furthermore, to predict the total number of cases globally , the authors Gothai, E. et al have used a machine learning approach [11]. The data set was collected from the repository of John Hopkins University, data collected had information regarding the active cases, deaths caused by coronavirus and day-to-day new cases that were being reported in the world. Dataset consisted of data from January 2020 to December 2020. Initially the data was preprocessed to converted into standardized data which had dates along with multiple other attributes such as country, province, serial num, date of observation and total number of patients recovered from coronavirus. In total 8 attributes were available in the dataset used for which an open access was given by World Health Organization (WHO) to all the hospitals and researchers. To refine the results unnecessary text was removed by the author moreover, they used lemmatization along with punctuation to prepare the dataset for use in machine learning. To increase the distinction accuracy and classification, stop terms and icons were excluded. Feature extraction was performed by taking out numerous features from preprocessed clinical reports which were then converted to probabilistic numbers according to the sementics. Pandas and NumPy library was used to identify relevant features for the study. Relevant features which were extracted for the study included recovered cases, death and confirmed cases etc. classification was achieved by extraction of relevant features and weightages were assigned to the features after which same inputs were provided to the ML algorithm. To predict the total number of confirmed cases in the world multiple algorithms were employed such as Linear regression, Support Vector Regression (SVR) and Holt-winter model for time series forecasting. Model was trained for all algorithms using same dataset by providing relevant features as training input. The proposed methodology included; preprocessing of data, feature extraction, and then machine learning models were applied. Study showed that in Linear Regression the data was classified into 4 types i.e. closed cases, death cases, confirmed cases and active cases. Same classification was used for Support Vector Machine (SVM) and results were compared. It was noted that large difference between obtained results and real time dataset was there, so Time series algorithm was used to train the model. To obtain more precise predictions exponential smoothing method was used and results showed that Holt's winter linear model can predict with an accuracy of 87% as compared to other two models. Accuracy of Holt's winter model is provided in comparison to the other two models used and proper model evaluation is not provided based on R2, RSME, MSE or MAE etc. Author also mentioned the use of Python libraries which were used to reflect and measure the real time trends of coronavirus in the world by using graphs and the curve base on the virus's trend month wise.

In the paper [12], two models are proposed for the prediction of covid-19, so that effective measures can be taken accordingly. First, a mathematical model is used to identify the significance of various parameters on the total number of active cases and predictions can be made for number of cases for future. Second, the Fourier decomposition method was used for the trend in reported cases on a daily basis. Fourier decomposition is a widely used method for analysis of different physical phenomena. In this method the time series was decomposed into cosine and sine basis functions. Coronavirus time-series was decomposed into frequency bands and various trends along with variabilities were obtained. The trends obtained were then fitted with a Gaussian mixture functions to predict the scale of epidemic. Lastly, the Bi-modal Gaussian mixture model is used for speculating the cases. The prediction is made with a 95% confidence interval, which demonstrates a promising outcome.

In order to foresee the cases , specifically in India , the authors Anuradha .T et al have proposed a methodology that estimates 30 days ahead[13]. The data used in this paper is from 30th January 2020 to 4th April 2020. In this paper, methods like long short-term memory methods are used with 80% training of data and 20% for testing. Furthermore, the parameters being predicted are; total positive cases, total recovered cases, etc. Lastly, the analysis has also been done on the effects of lockdown, the impact of transformation, and social isolation.

The authors Mohsen Maleki et al [14], have put forth a methodology for the prediction of confirmed and recovered number of cases. The suggested methodology considers autoregresive time series models i.e two-piece scale mixture models including asymetric heavy-tailed non-Gausian time series and symmetric Gausian models. This model identifies the probabilistic behaviour of the real time values taking in account the linear combination and pattern of recent values. The author initially tested various autoregressive models which were proposed. These models were fitted to the past data of recovered and confirmed cases around the globe. Autoregressive time series model was selected as this was the best fit for all datasets. Data for Jan 2020 till April 2020 was taken into account and TP-SMN-AR time series was fitted. Proposed time series plots were not found to be stationary due to the increasing number which reflects a trend. To obtain a stationary data suitable transformations were applied . Afterwards the using selection criteria of model the autoregressive models were selected and fitted to the stationary data containing recovered and confirmed cases. Estimated errors which are also known as residuals were reviewed which showed appropriate performance of model with respect to the stationary series representing recovered and confirmed cases in dataset. To analyse the suitability of the model autocorrelation function plots of residual were also reviewed which provided a promising result of training. To test the model 10 records from head of the data were omitted and model was fitted to get the predictions which gave 98% confidence intervals for the analysis. For model evaluation and determine the accuracy Mean Relative Percentage Error (MAPE) was used. For recovered cases the value of MAPE was 1.6% whereas, for recovered cases of coronavirus it was 0.22%. To support the claim that proposed model is more reasonable than other prediction models author discussed the model selection criteria such as Bayesian Information Criteria (BIC), Akaike Information Criteria (AIC) and Box-Pierce test on residuals. Lastly, according to the results, the model had 98 % confidence intervals, which is commendable considering the availability of the data moreover, only 2 features were selected randomly, no feature selection is performed and handling of missing values is addressed in this study.

Similarly, in efforts to foresee recovered, confirmed, and death cases, the author Tandon, H. et al have proposed an Auto-Regressive Integrated Moving Average model (ARIMA) [15]. The dataset was collected from the repository of John Hopkins University from January 22^{nd,} 2020 to April 13th 2020. ARIMA was initially applied to the time-series data of confirmed coronavirus cases afterwards ACF graph and Partial autocorrelation (PACF) graph is obtained to identify initial number of ARIMA models. The variance in stationary and normality was then tested. To check the accuracy of the model MAPE, MSD and MAD values are determined. These values represented that which model is most appropriate for for predictions. Once the model is selected and fitted, its parameters were projected. After performing the comparative analysis it was observed that the outcome of the ARIMA model was more accurate as compared to Linear Trend, Single, and Double Exponential, Quadratic Linear, Moving Average, and S-curve Trend. The performance metrics MAPE, MAD, and MSD of the ARIMA (2,2,2) model were 4.1, 58.3, 25319.5, respectively, which were least values among other models. Lastly, the proposed model has the potential to help the government take preventive measures in accordance with the upcoming situation.

In the paper [16], the authors have compared different deep learning-based models for the prediction of total individuals who may get infected with Coronavirus 2019. The dataset is taken from the Ministry of Health and Family Welfare, India. In addition, it has 32 individual time-series data of confirmed cases. Moreover, the training models used are Stacked LSTM, Bi-directional LSTM, and Convolutional LSTM. Authors initially divided the target region into three different groups based on total number of positive cases and daily rise. Three zones were: 1: Mild Zone- areas where number of positive cases are less than 200 and day-to-day repoted cases rise is below 2%. 2: Moderate Zone: regions where number of positive patients is between 200 to 2000 and day-to-day increment is less than 5%. 3: Severe Zone: regions with positive patients are above 2000 and day-to-day rise in cases is greater than 5%. Analysis is performed using open source libraries such as Numpy, Pandas, Python as ageneral purpose programming language. Multiple models were used in this study to understand and learn the dynamic dependent structure in the data and also to identify the learning sequence present in the data with respect to any specific region. Data is provided to these models containing historic data. Hyper-parameter tuning of each of these models is done thoroughly and parameters are selected. Adam optimizer is used for optimizing the mean squared error loss. After calculation of error loss the optimal model is selected i.e. Bi-LSTM. Bi-LSTM has limited error range with minimum average error among all models and is more suitable for forecasting purposes as compared to other models. After the hyper-parameter tuning, the best model is selected for the prediction. The Bi-directional LSTM gave the most accurate results with an error of 3%. Even though the proposed model can give an effective prediction, the deep learning models require more data to perform well, which was not encountered in this paper.

In the paper [17], the authors have proposed methods for diagnosis and forecasting of confirming cases. Three models for predictions are compared: Prophet Algorithm, Auto-regressive integrated moving average model, and long short-term memory neural network. Furthermore, the comparison of the model was based on the evaluation metrics; correlation coefficient, accuracy, and RMSE. However, the results were the most accurate for prophet algorithms. The accuracy of confirmed, recovered and death cases, based on Jordan were: 97.08%, 79.39%, and 86.82%, respectively. Similarly, the aforementioned cases were also analyzed for countries as well.

The paper [18] focuses on forecasting of covid-19 based on Brazil, for one, three, and six days ahead. The evaluated models for prediction are; support vector regression, ridge regression, random forest, cubist regression, autoregressive integrated moving average, and stacking- ensemble learning. The dataset is obtained from Brazilian state health offices. To understand data author used visualizations such as graphs and heat maps. The data obtained was split in to test and train data. Six last observations were taken as test data whereas, all other observations were used to train the models. To handle the skewness, training data is centered by its mean and divided by the standard deviation. Recursive strategy is used to forecast next day's expected cases. The evaluation is based on improvement index, symmetric mean absolute percentage error, and mean absolute error. The best performing model among them was Support vector regression, and the worse performing model was Random forest. Overall, the error range of the models is; 1.02%–5.63%, 0.87%–3.51%, and 0.95%–6.90% in three, one, and six-days-ahead, respectively. Nevertheless, the paper is based on the cases in Brazil only, the models should be more general so that the prediction can be made globally.

In paper [20] segmented Poisson model is employed by the authors Xiaolei Zhang et al to analyze the new corona virus cases, day-to-day outbreaks in western countries of G-7. Authors performed a statistical prediction to determine the period when daily new cases are at peak and how long does the outbreak lasts, moreover, the percentage of population that will get infected by coronavirus. They used Wind database to get the data for daily new confirmed cases of coronavirus for G-7 countries. To predict the spread of coronavirus government enforcements such as lockdown and stay at home policies were also accounted for in this study. Power las was combined with exponential law to get daily new cases based on segmented poission model. By performing statistical analysis the parameter estimate with confidence intervals of 95% for all G-7 countries. A 14 day prediction model for daily new cases was developed which showed the infection rate of 0.28% based on the final size as per the data used. The author claimed that combination of exponential law with segmented Poission model to forecast the new daily cases of coronavirus observed and estimated cases are in good agreement and predictions are based on assumptions that government interventions will remain unchanged till the estimated end dates.

The paper [21] focuses on the prediction of the final size for Coronavirus-2019 using machine learning. The authors Lamiaa A. Amar et al explained that to predict the final size for the Covid-19 regressing models are used and dataset from Our World in Data (OWID) is used but specifically for Egypt. Multiple types of regression model are used in this study such as Exponential Regression model, Polynomial regression model and logit growth regression model. Data transformation and feature selection was not performed and models were applied directly to the data set. While performing regression analysis correlation coefficients were calculated which represent the intensity of linear relationship between variables. The values range between -1.0 and 1.0. To check the goodness of fit authors calculated both simple R2 and adjusted R2 and according to results provided Logit regression model had the highest 0.99 R2 value. Authors did not preprocess the data to standardize it and did not perform feature engineering on the dataset. Ambiguities and anomalies might be present in dataset which may affect the prediction results.

2.3 Critical Analysis

During the literature review it was discerned that no work has been done regarding the feature selection and engineering in any of the reviewed papers. Using data preprocessing, you can make your Machine Learning process more efficient and streamlined. Moreover, using machine learning algorithms the preprocessing of data is a very crucial step to achieve accurate and optimal outcomes. The data was separated into categorical and numerical features to tackle their missing values accordingly. Another gap that was identified was that none of the work reviewed was handling the skewness in the data. Our model applied different transformation techniques to highlight the skewness present in the data and uses yoe johnson’s technique to normally distribute the data. Moreover, min-max feature scaling is being implemented in our model to scale the features on a specific range which is yet another gap identified in the work reviewed. Our model also makes use of feature selection which is one of the ambiguities highlighted in the reviewed papers as feature selection helps in the faster execution of the model. Out of 113 attributes, 17 attributes are used to build a regression model for the prediction of the virus after which we performed model fitting to achieve optimal results. Furthermore, most of the reviewed work makes predictions based on specific countries whereas the proposed model is making generalized predictions taking various continents and locations into consideration. In conclusion, most of the reviewed work has applied models directly to the data collected and aren't generalized i.e they predict the wave based on a specific country. Datasets used by other authors had insufficient data as it was either specific to a location or max 1 month data was taken from the dataset. It is always recommended that data repository should have enough data to run into huge numbers, problem caused by insufficient data is that it increases the variance, variance can be arbitrarily defined as the final value given to a particular parameter (time) such as performance gain or indicates how the data is spread. The comparative analysis of the reviewed work is given in table 1. This table indicates the methodology deployed in our model versus the methodology utilised by the work of others. It can be clearly seen that our model employs all these techniques for a better and efficient system.

Gaps	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	Our model
Handling missing values	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
Data Transformation	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Feature Extraction	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Feature preprocessing	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓
Feature Scaling	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Feature Selection	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
General model	✗	✗	✗	✓	✗	✓	✗	✗	✓	✗	✓

Table 1 "Critical Analysis"

Chapter 3

Proposed Methodology & Implementation

CHAPTER 3: PROPOSED METHODOLOGY & IMPLEMENTATION

3.1 Problem Definition

Pandemics have occurred throughout time, taking millions of lives worldwide. The first case of the novel coronavirus whose origins trace back to Wuhan city, China was first reported in 2019. Since the virus has prevailed globally,the situation has been quite difficult for the contemporary world. A large number of lives have been influenced either straightforwardly from the infection or in a roundabout way from the insurances governments have needed to take to control the infection. Moderate the infection however much as could reasonably be expected so that lives can be saved and the influenced individuals can observe normalcy as usual once more. The coronavirus is a virus that can cause illnesses in the respiratory system of the human body. Common symptoms for coronavirus are fever or chills, cough, shortness of breath or difficulty breathing, tiredness, muscle or body aches, headaches, congestion or runny nose, sore throat, nausea or vomiting and diarrhea, new loss of taste or smell,. This diseases anyone of any age. The youngest covid patient who contracted this deadly virus was 4 weeks of age.

Developments in technology are causing tremendous changes in many life sectors, may it be education, work, industry, or transportation. In the domain of health, computerization is still a work in progress. A lot of work has been done to develop such tools that may help with dealing with various diseases, for example, virtual reality simulations are used to help people overcome their phobias. With digitalization upsetting each industry, including medical services, the capacity to catch, share and convey information is turning into a high need. ML, big data and computerized reasoning (AI) can assist with tending to the difficulties that tremendous measures of information present. ML and AI can likewise help medical services associations fulfill developing clinical needs, further develop activities and lower costs. Moreover , machine learning advancements can help medical services professionals distinguish and treat sickness more effectively and with more accuracy and customized care.

As the humongous amounts of data is being generated, there is a need to effectively process it by saving a strenuous amount of time and work. Data is available on the internet in various formats and is easily accessible; however, to manipulate it and to derive a logical explanation from it takes a lot of human effort. To minimise this human effort in order to study the trends of the past and the future a predictive approach can be applied by using algorithms of machine learning. In machine learning different algorithms can be applied to reach an optimal solution. The process of manipulating the data is very transparent. Large amounts of data with millions of data points can be easily visualized and interpreted. Machine learning and IOT can be used to mitigate the covid-19. ML generated models can help in battling any upcoming pandemic. We can reuse the existing models with slight variation and train them on new types of datasets to mitigate the spread of any future pandemic. Moreover, the performance of these ML models can be improved by incorporating new machine learning methods. The performance gaps can be filled by revising the models to incorporate new advancements in machine learning. This work aims to develop a regression model for prediction of the next wave of coronavirus diseases using covid-19 dataset from Our World in Data as a data source. The goal of this research is to actively predict the next wave of the deadly virus across the globe.

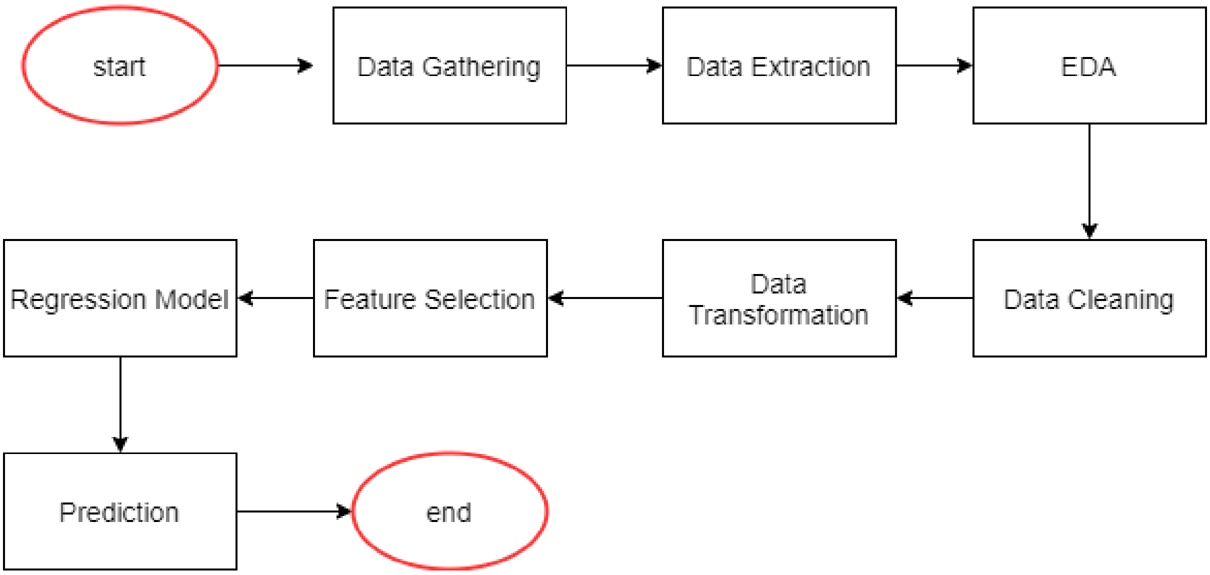


Figure 1 "Flow chart of the methodology"

The figure above shows the steps that should be followed in order to create a regression model. Firstly the data is gathered into a database or data source. It is then extracted from the source it was stored in. Then exploratory data analysis is performed on the gathered data to determine the gaps and outliers followed by data cleaning. Moreover, the data is transformed in order to achieve optimal results. Then certain features are selected to achieve higher accuracy and an optimal model. Lastly the regression model is applied on the selected data which results in deriving predictions.

3.2 Research Methodology

The following steps are carried out in order to extract data from owid repository and reading it into the notebook

3.2.1 Data gathering, extraction and loading

As mentioned above we choose Our World in Data as data source to create the regression model. Our World in Data is an open platform providing free services. The data resources are accessible to researchers, professionals of healthcare and common man. The data is gathered and collected based on the work of global community scholars and researchers. The data is easily accessible and can be downloaded indisputably from github. The data is extracted from the github repository and converted to a comma separated value (CSV) file. The downloaded data file is of 24.9 MB's after extracting it from the .rar folder. The process can be seen in fig. 2 & 3. The steps are given below:

1. Download zipped file of the covid-19 dataset from Our World in Data repository.

2. Extract the file using win.rar.

3. Select the owid-covid-data.csv and place it in the same folder as your python files

4. Upload the data file in jupyter.

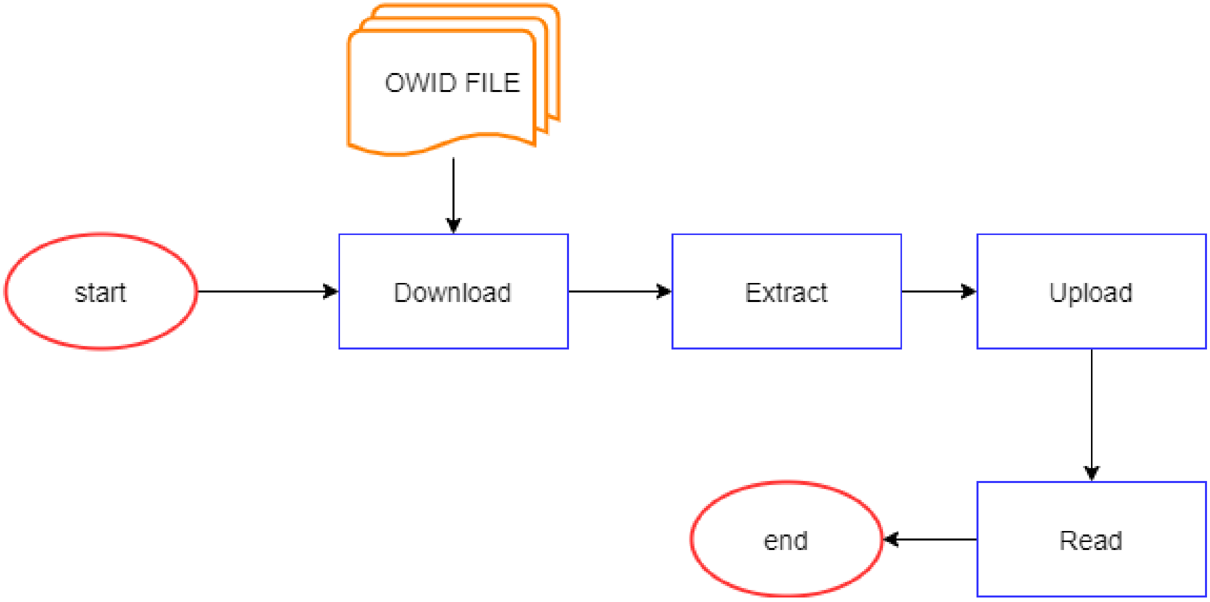
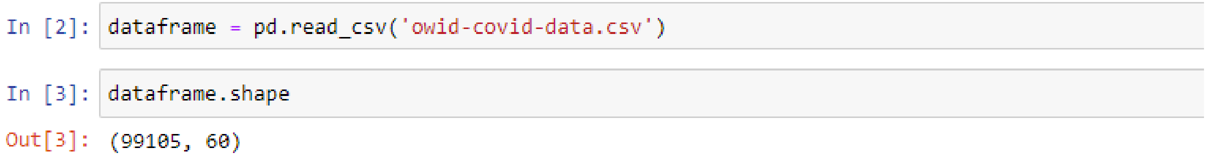


Figure 2 "Process of extracting and loading data"

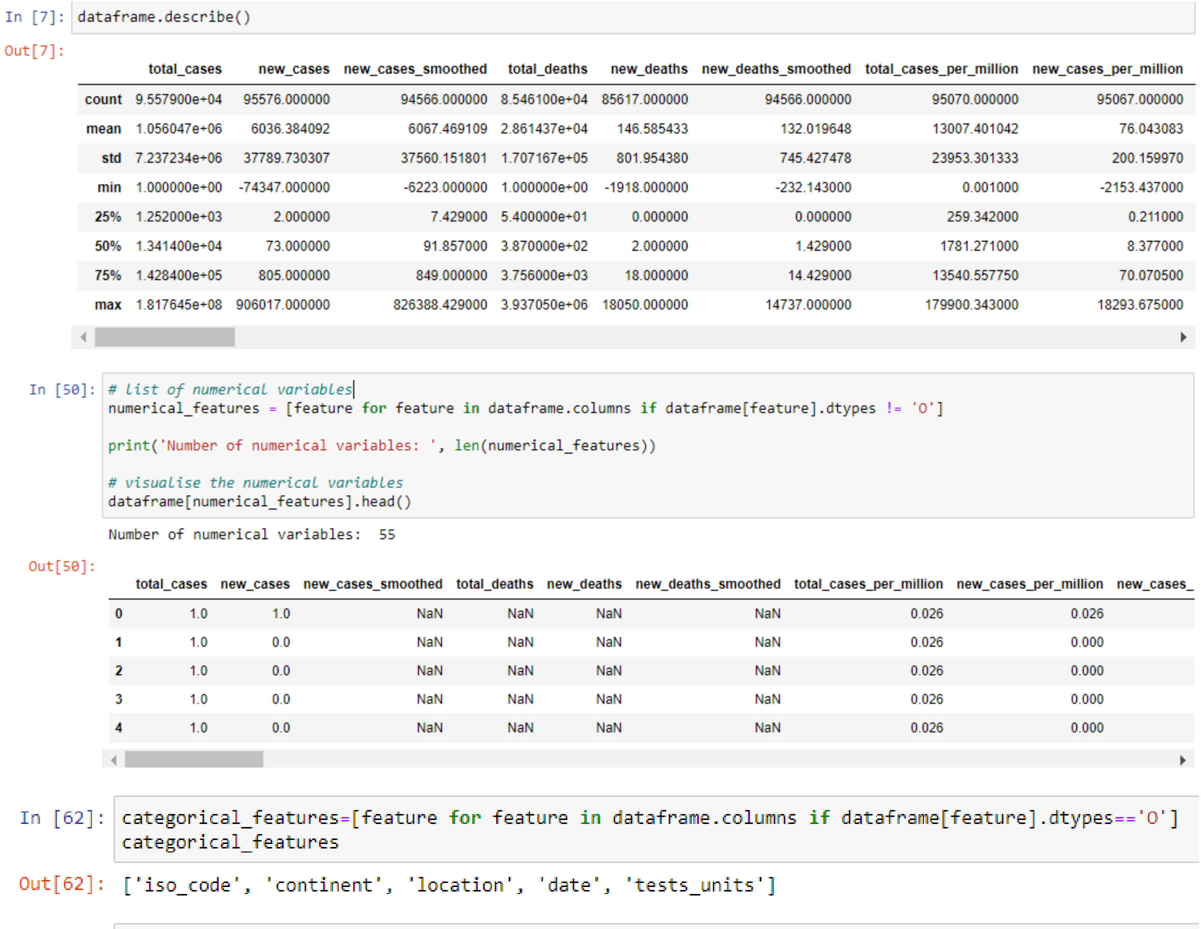


```
In [2]:  dataframe = pd.read_csv('owid-covid-data.csv')

In [3]:  dataframe.shape
Out[3]:  (99105, 60)
```

Figure 3 “Process of reading data”

3.2.2 Data Analysis

Exploratory Data Analysis (EDA) is applied in order to dissect and research informational indexes and sum up their fundamental attributes, regularly utilizing information representation techniques. EDA is essentially used to perceive what information can uncover past the conventional demonstrating or theory testing task and gives a superior comprehension of informational index factors and the connections between them [10]. EDA is applied to the owid covid-19 dataset to unravel hidden information, It can be seen in figure. 4.



```
In [7]:  dataframe.describe()
Out[7]:
```

	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_per_million
count	9.557900e+04	95576.000000	94566.000000	8.546100e+04	85617.000000	94566.000000	95070.000000	95067.000000
mean	1.056047e+06	6036.384092	6067.469109	2.861437e+04	146.585433	132.019648	13007.401042	76.043083
std	7.237234e+06	37789.730307	37560.151801	1.707167e+05	801.954380	745.427478	23953.301333	200.159970
min	1.000000e+00	-74347.000000	-6223.000000	1.000000e+00	-1918.000000	-232.143000	0.001000	-2153.437000
25%	1.252000e+03	2.000000	7.429000	5.400000e+01	0.000000	0.000000	259.342000	0.211000
50%	1.341400e+04	73.000000	91.857000	3.870000e+02	2.000000	1.429000	1781.271000	8.377000
75%	1.428400e+05	805.000000	849.000000	3.756000e+03	18.000000	14.429000	13540.557750	70.070500
max	1.817645e+08	906017.000000	826388.429000	3.937050e+06	18050.000000	14737.000000	179900.343000	18293.675000

```
In [50]: # list of numerical variables
         numerical_features = [feature for feature in dataframe.columns if dataframe[feature].dtypes != 'O']

         print('Number of numerical variables: ', len(numerical_features))

         # visualise the numerical variables
         dataframe[numerical_features].head()

         Number of numerical variables:  55
Out[50]:
```

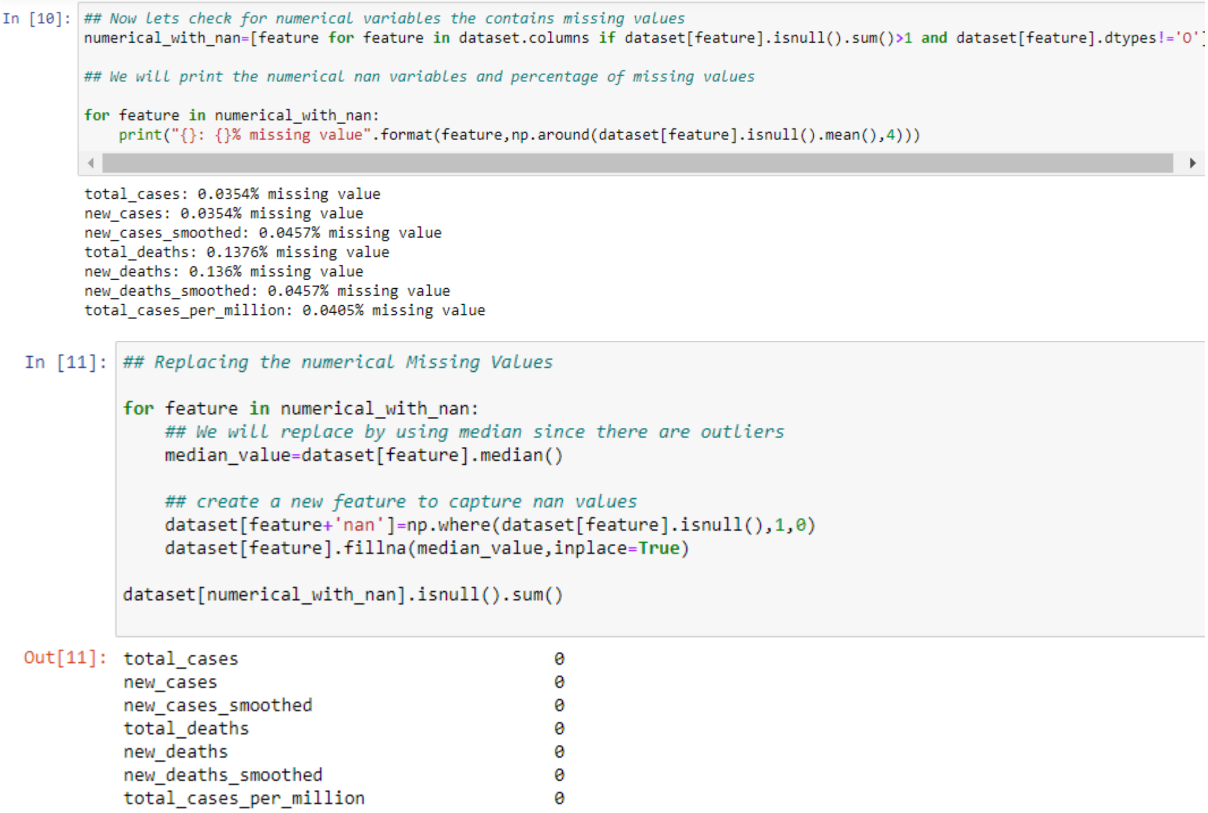
	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_per_million	new_cases_
0	1.0	1.0	NaN	NaN	NaN	NaN	0.026	0.026	
1	1.0	0.0	NaN	NaN	NaN	NaN	0.026	0.000	
2	1.0	0.0	NaN	NaN	NaN	NaN	0.026	0.000	
3	1.0	0.0	NaN	NaN	NaN	NaN	0.026	0.000	
4	1.0	0.0	NaN	NaN	NaN	NaN	0.026	0.000	

```
In [62]: categorical_features=[feature for feature in dataframe.columns if dataframe[feature].dtypes=='O']
         categorical_features
Out[62]: ['iso_code', 'continent', 'location', 'date', 'tests_units']
```

Figure 4 “Exploratory Data Analysis”

3.2.3 Data Cleaning

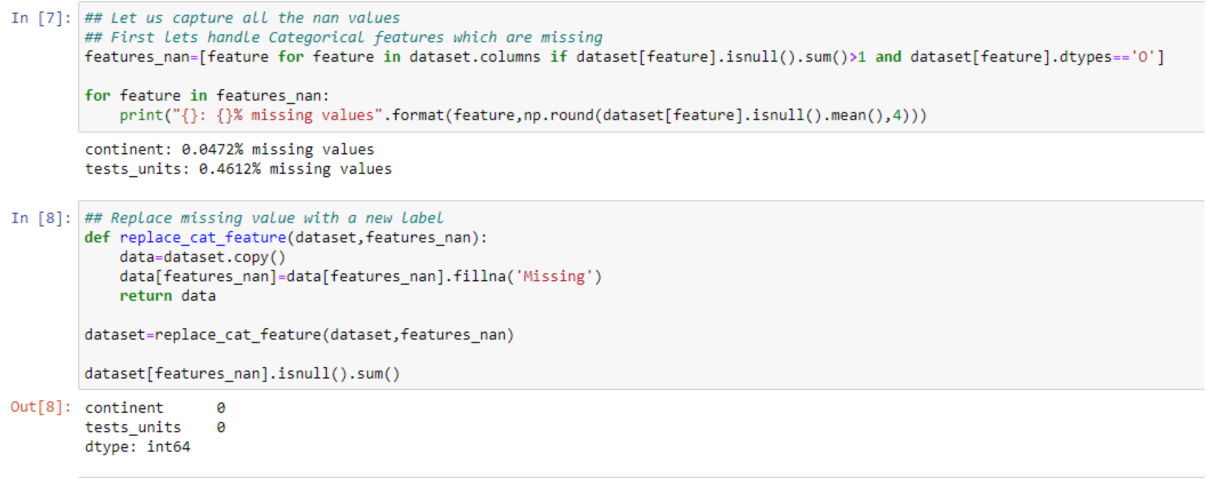
The dataset extracted from OWID has huge amounts of numerical and categorical missing values. In order to achieve an optimal solution we need to handle the missing values and replace them with suitable techniques. It can be seen in figure. 5 & 6.



```
In [10]: ## Now lets check for numerical variables the contains missing values
         numerical_with_nan=[feature for feature in dataset.columns if dataset[feature].isnull().sum()>1 and dataset[feature].dtypes!='O'

         ## We will print the numerical nan variables and percentage of missing values

         for feature in numerical_with_nan:
             print("{}: {}% missing value".format(feature,np.around(dataset[feature].isnull().mean(),4)))

         total_cases: 0.0354% missing value
         new_cases: 0.0354% missing value
         new_cases_smoothed: 0.0457% missing value
         total_deaths: 0.1376% missing value
         new_deaths: 0.136% missing value
         new_deaths_smoothed: 0.0457% missing value
         total_cases_per_million: 0.0405% missing value
```

```
In [11]: ## Replacing the numerical Missing Values

         for feature in numerical_with_nan:
             ## We will replace by using median since there are outliers
             median_value=dataset[feature].median()

             ## create a new feature to capture nan values
             dataset[feature+'nan']=np.where(dataset[feature].isnull(),1,0)
             dataset[feature].fillna(median_value,inplace=True)

         dataset[numerical_with_nan].isnull().sum()

Out[11]: total_cases                              0
         new_cases                                0
         new_cases_smoothed                       0
         total_deaths                             0
         new_deaths                               0
         new_deaths_smoothed                      0
         total_cases_per_million                  0
```

Figure 5 "Process of cleaning numerical data"



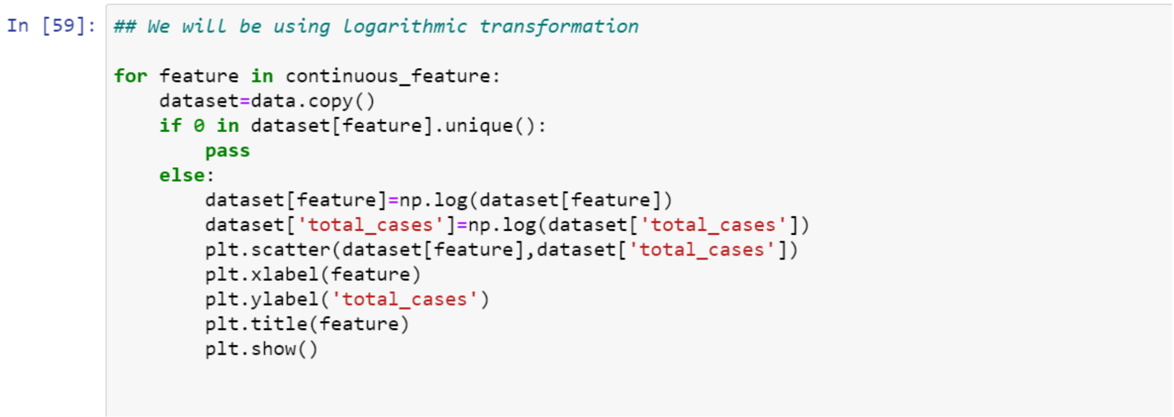
```
In [7]: ## Let us capture all the nan values
        ## First lets handle Categorical features which are missing
        features_nan=[feature for feature in dataset.columns if dataset[feature].isnull().sum()>1 and dataset[feature].dtypes=='O']

        for feature in features_nan:
            print("{}: {}% missing values".format(feature,np.round(dataset[feature].isnull().mean(),4)))

        continent: 0.0472% missing values
        tests_units: 0.4612% missing values
```

```
In [8]: ## Replace missing value with a new label
        def replace_cat_feature(dataset,features_nan):
            data=dataset.copy()
            data[features_nan]=data[features_nan].fillna('Missing')
            return data

        dataset=replace_cat_feature(dataset,features_nan)

        dataset[features_nan].isnull().sum()

Out[8]: continent      0
        tests_units    0
        dtype: int64
```

Figure 6 "Process of cleaning categorical data"

3.2.4 Data Transformation

Transformation and scaling techniques are applied to numerical and categorical features in the data. It is shown in figure. 7 and 8. After data cleansing, there was a dire need of data transformation as original data was skewed and missing values were replaced in prior stages. To transform the OWID dataset in order to reduce skewness in raw data we used Joe-Jhonson transformation which can handle both negative values as well as 0 values.

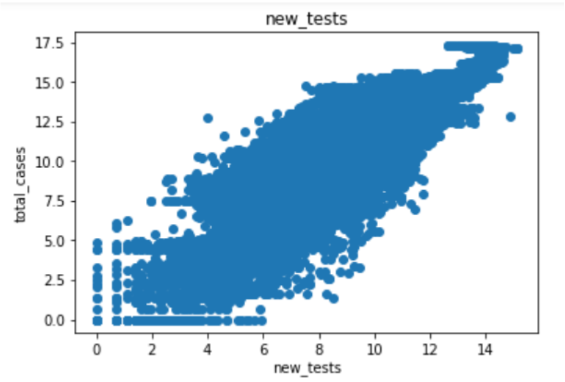


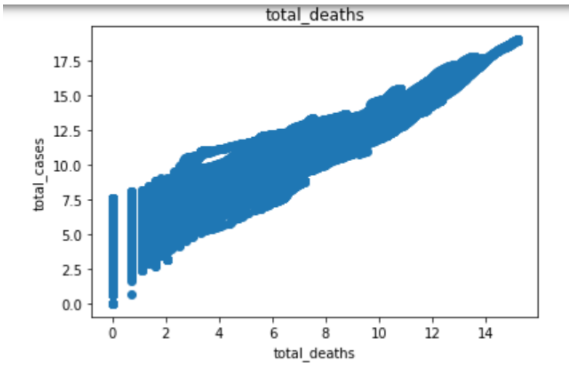
In [59]:

```
## We will be using logarithmic transformation

for feature in continuous_feature:
    dataset=data.copy()
    if 0 in dataset[feature].unique():
        pass
    else:
        dataset[feature]=np.log(dataset[feature])
        dataset['total_cases']=np.log(dataset['total_cases'])
        plt.scatter(dataset[feature],dataset['total_cases'])
        plt.xlabel(feature)
        plt.ylabel('total_cases')
        plt.title(feature)
        plt.show()
```

Figure 7.1 “Logarithmic Transformation”

Scatter plots are analyzed to get a better understanding of the data and identify if any outliers are present in the dataset. Figure 7.2. below shows set of scatter plots for total deaths, new cases reported world wide, total number of cases and hospital beds per thousand people. These plots show the relationship of two variables which helps is determining the pattern and identifying if any variable is dependent or independent.





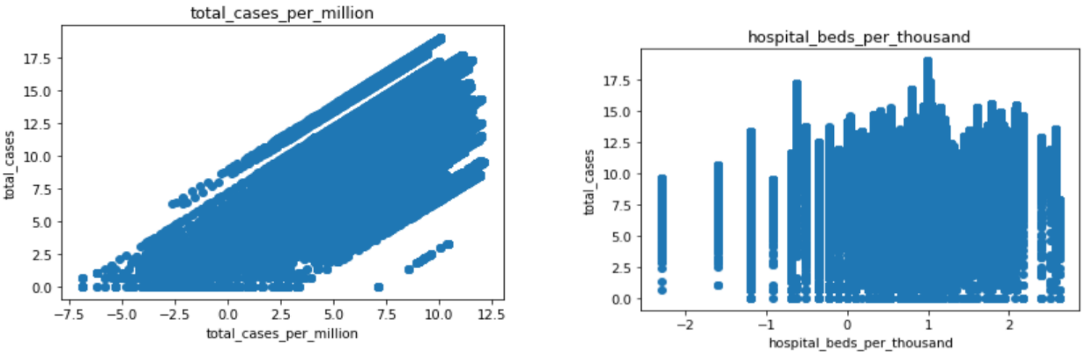
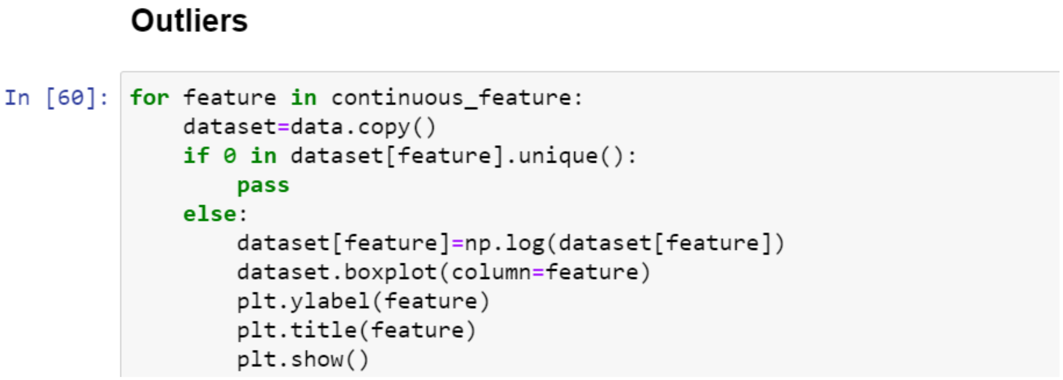


Figure 7.2 "Scatter Plots"

i. Outliers

The outliers are identified and handled while before transformation of data. Outliers can be due to multiple reasons such as computer glitch or human error. Outliers tend to manipulate the mean value which negatively impacts the prediction accuracy. Figure 7.3 shows the box plot for outliers.



Outliers

```
In [60]: for feature in continuous_feature:
             dataset=data.copy()
             if 0 in dataset[feature].unique():
                 pass
             else:
                 dataset[feature]=np.log(dataset[feature])
                 dataset.boxplot(column=feature)
                 plt.ylabel(feature)
                 plt.title(feature)
                 plt.show()
```

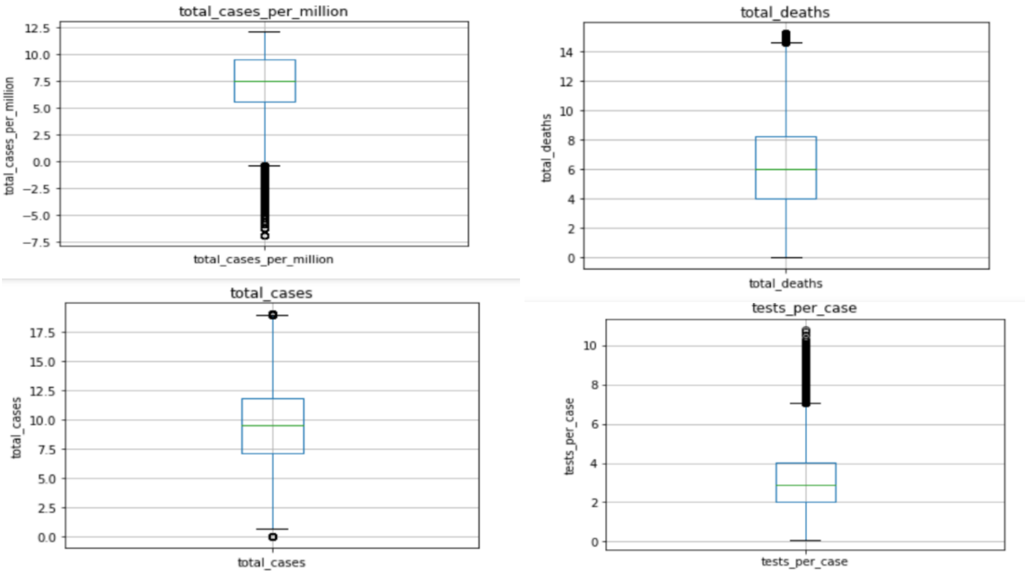
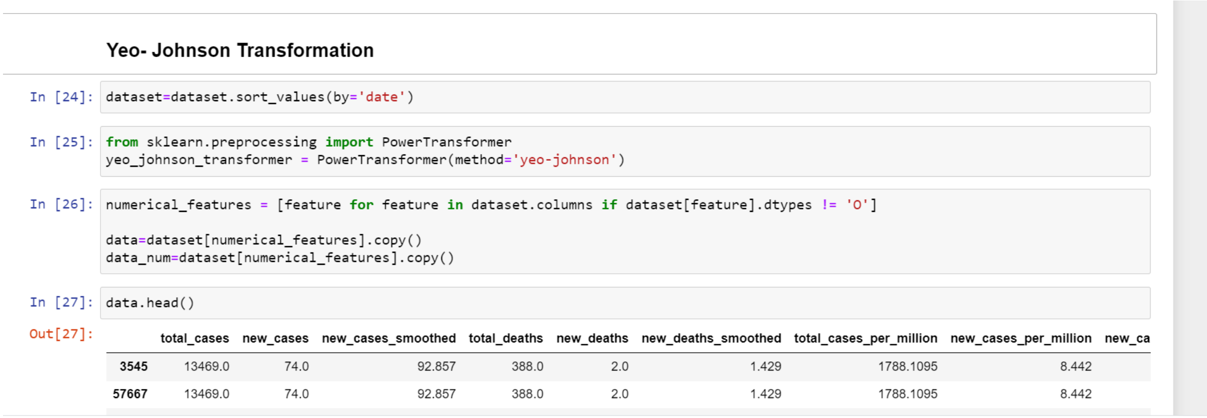


Figure 7.3 "Outliers"

ii. Joe-Jhonson Transformation

The Yeo-Johnson transformation is not much different from the Box-Cox. The difference is that it does not require the input variable quantity to be positive. Data is normally distributed and skewness is reduced which enhances the chances for better predictions. This transformation is classically performed targeting the outcome variable for which residuals are used for a statistical model. The outcome of this transformation is making the distributions more symmetric.



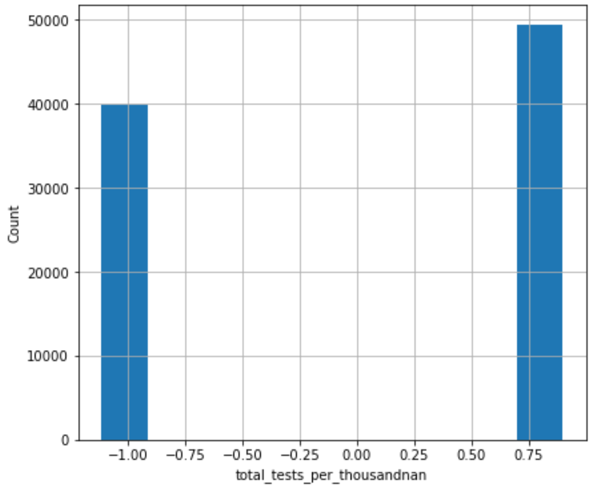
Yeo- Johnson Transformation

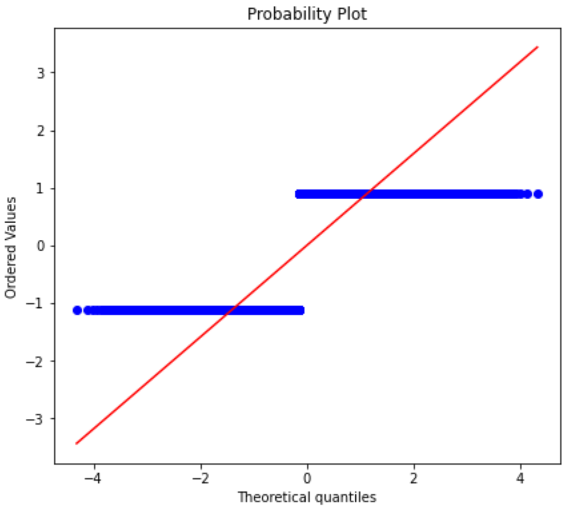
```
In [24]: dataset=dataset.sort_values(by='date')

In [25]: from sklearn.preprocessing import PowerTransformer
         yeo_johnson_transformer = PowerTransformer(method='yeo-johnson')

In [26]: numerical_features = [feature for feature in dataset.columns if dataset[feature].dtypes != 'O']

         data=dataset[numerical_features].copy()
         data_num=dataset[numerical_features].copy()

In [27]: data.head()

Out[27]:
```

	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_per_million	new_ca
3545	13469.0	74.0	92.857	388.0	2.0	1.429	1788.1095	8.442	
57667	13469.0	74.0	92.857	388.0	2.0	1.429	1788.1095	8.442	

Figure 7.4 "Application of Yoe-Jhonson Transformation"

Below graphs in figure 7.5 & 7.6 show examples from the OWID dataset showing skewness and transformed data after application of Yoe-Jhonson transformation.





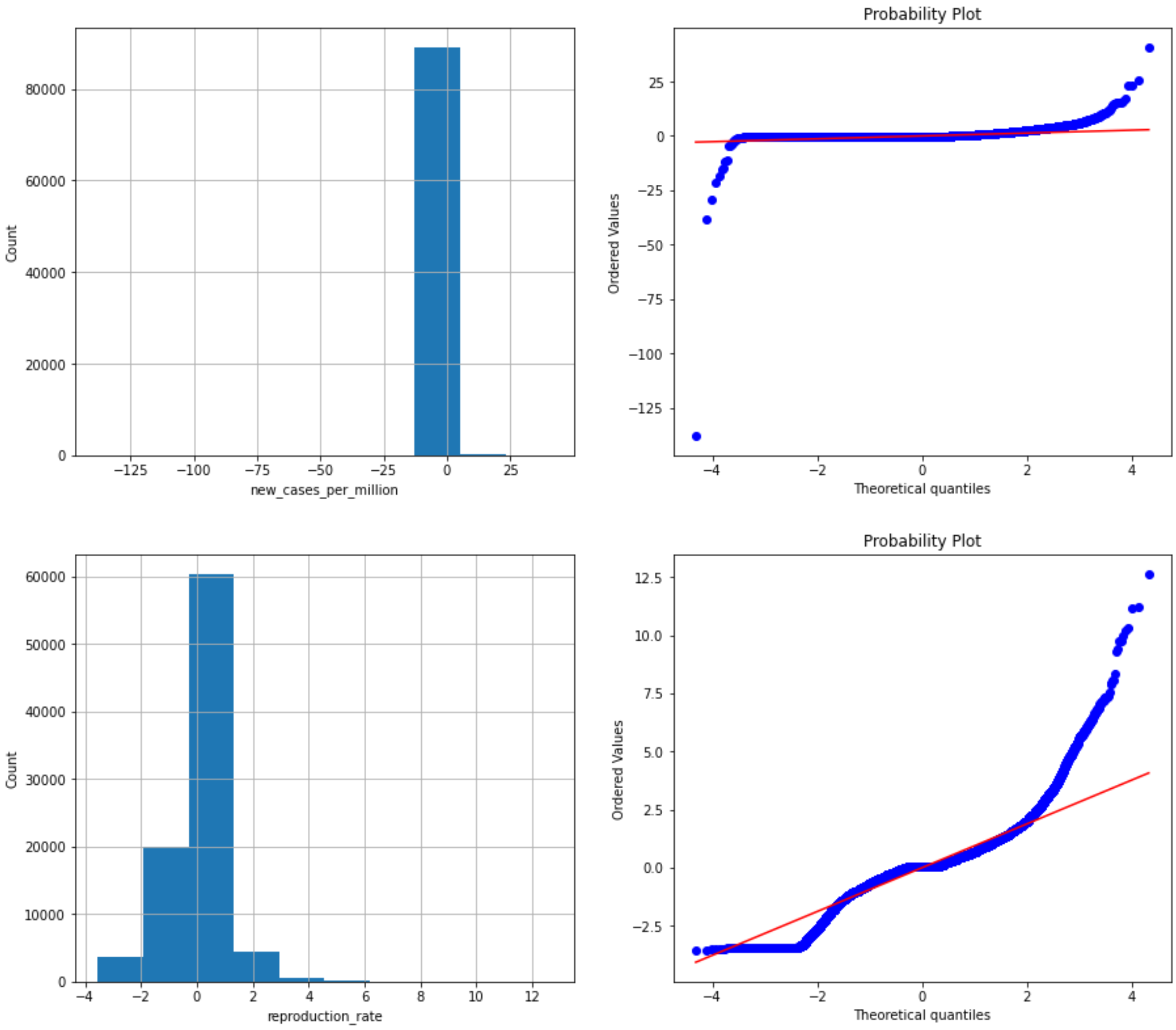


Figure 7.5 "Skewness of data"

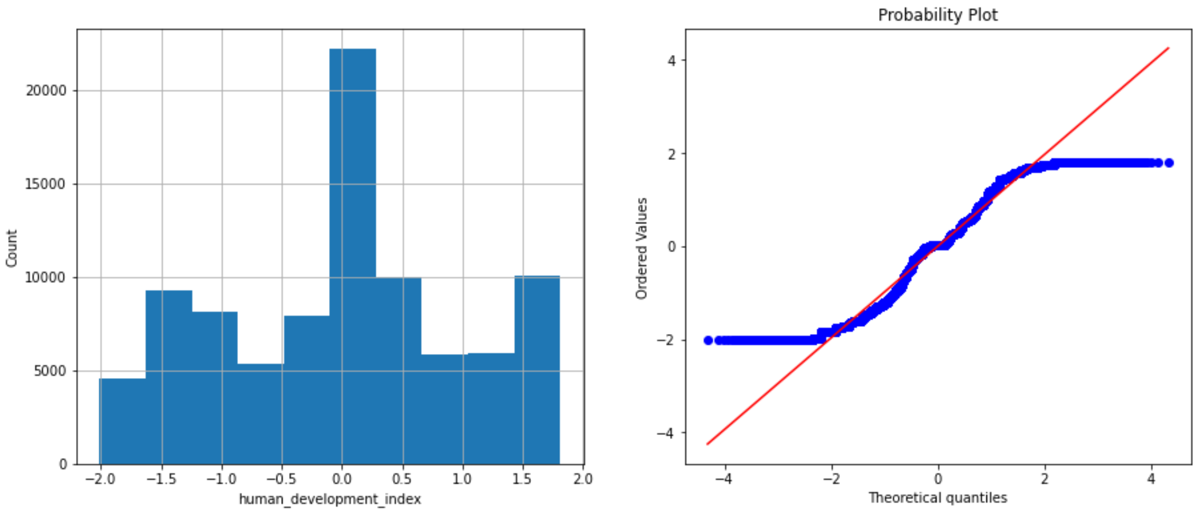


Figure 7.6 "Transformation applied on numerical features"

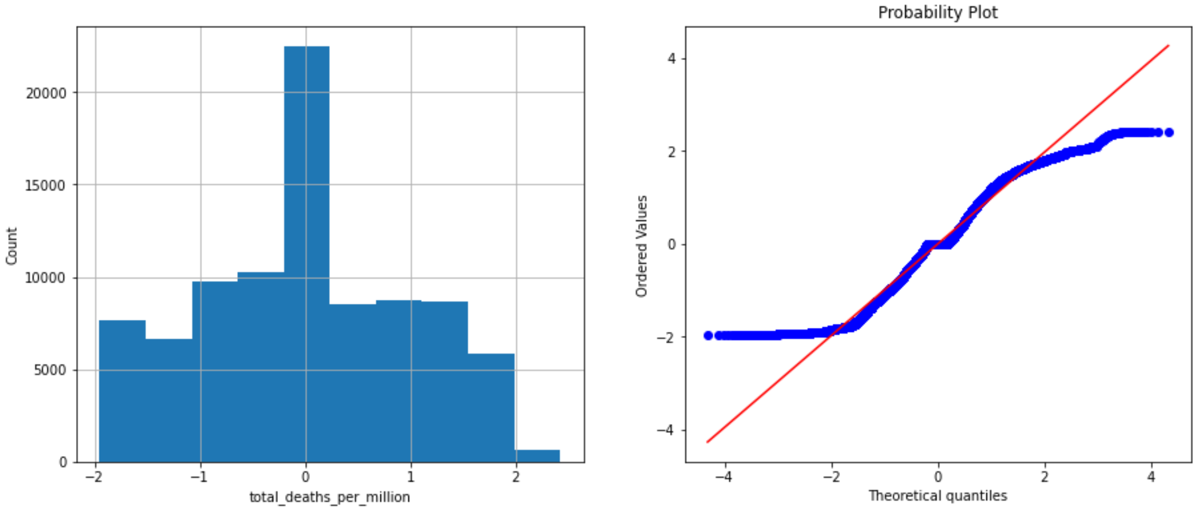


Figure 7.7 "Transformation applied total deaths per million"

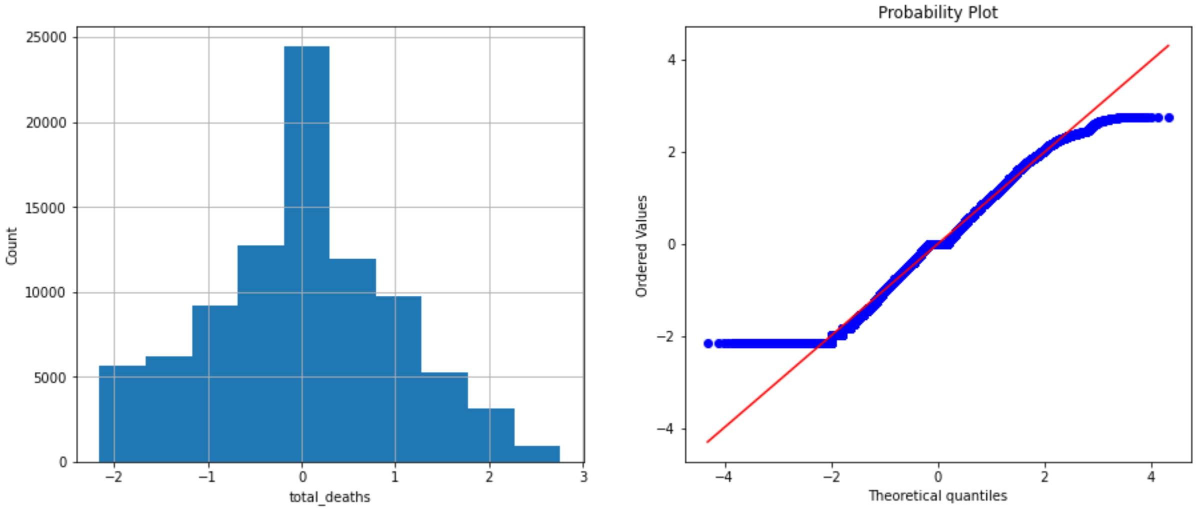


Figure 7.8 "Transformation applied on total deaths"

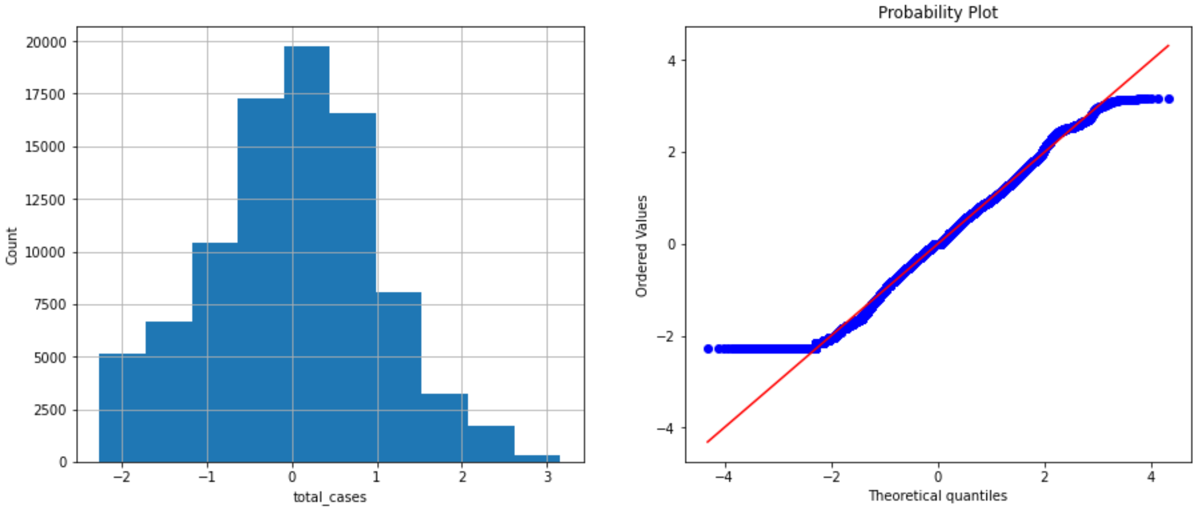
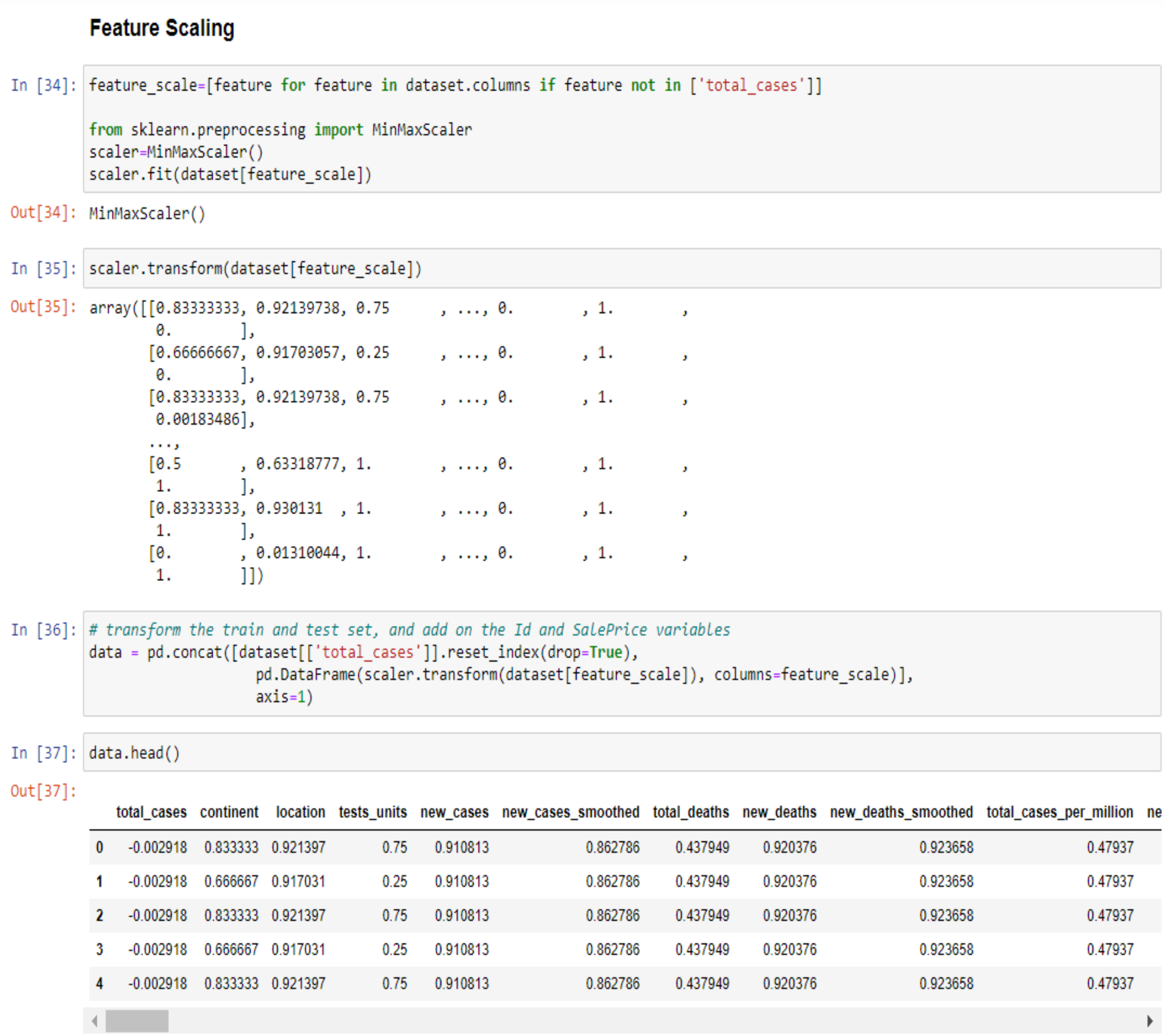


Figure 7.9 "Transformation applied on total cases"

3.2.5 Feature Scaling

In this study we have use normalization for feature scaling. Normalization is a method in which values are rescaled and shifted to ensure that they are ranging between 0 and 1. It is also known as Min-Max scaling.



Feature Scaling

In [34]:
```
feature_scale=[feature for feature in dataset.columns if feature not in ['total_cases']]

from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
scaler.fit(dataset[feature_scale])
```

Out[34]:
```
MinMaxScaler()
```

In [35]:
```
scaler.transform(dataset[feature_scale])
```

Out[35]:
```
array([[0.83333333, 0.92139738, 0.75      , ..., 0.        , 1.        ,
        0.        ],
       [0.66666667, 0.91703057, 0.25      , ..., 0.        , 1.        ,
        0.        ],
       [0.83333333, 0.92139738, 0.75      , ..., 0.        , 1.        ,
        0.00183486],
       ...,
       [0.5       , 0.63318777, 1.        , ..., 0.        , 1.        ,
        1.        ],
       [0.83333333, 0.930131  , 1.        , ..., 0.        , 1.        ,
        1.        ],
       [0.        , 0.01310044, 1.        , ..., 0.        , 1.        ,
        1.        ]])
```

In [36]:
```
# transform the train and test set, and add on the Id and SalePrice variables
data = pd.concat([dataset[['total_cases']].reset_index(drop=True),
                    pd.DataFrame(scaler.transform(dataset[feature_scale]), columns=feature_scale)],
                    axis=1)
```

In [37]:
```
data.head()
```

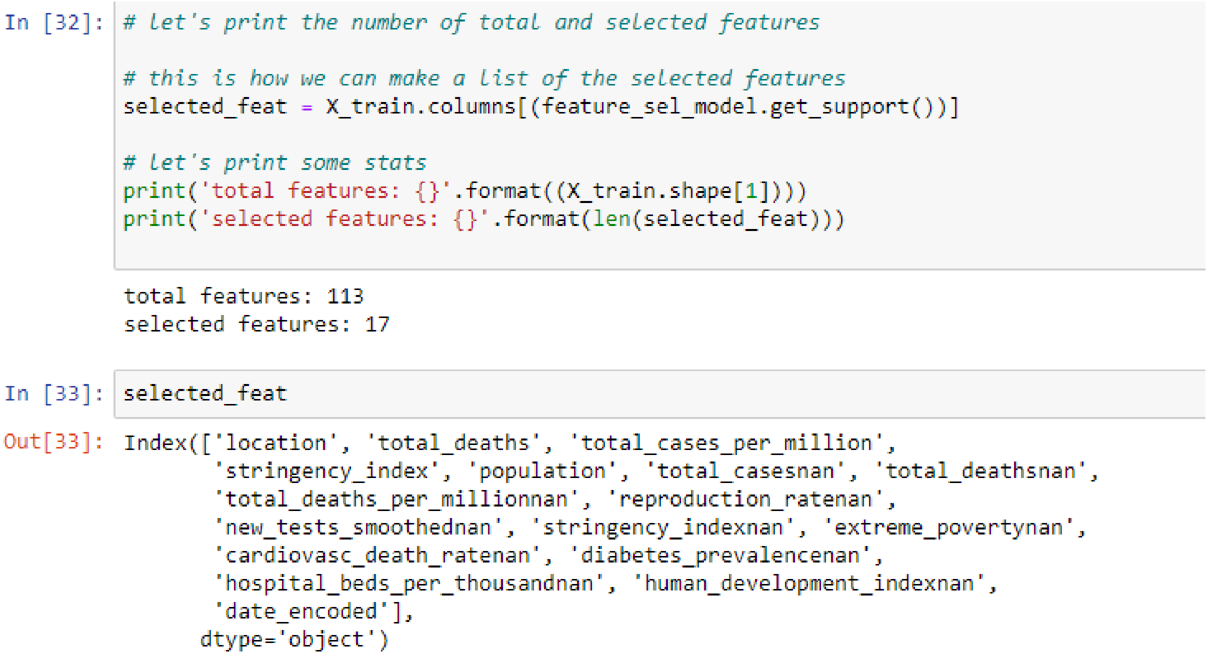
Out[37]:

	total_cases	continent	location	tests_units	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	ne
0	-0.002918	0.833333	0.921397	0.75	0.910813	0.862786	0.437949	0.920376	0.923658	0.47937	
1	-0.002918	0.666667	0.917031	0.25	0.910813	0.862786	0.437949	0.920376	0.923658	0.47937	
2	-0.002918	0.833333	0.921397	0.75	0.910813	0.862786	0.437949	0.920376	0.923658	0.47937	
3	-0.002918	0.666667	0.917031	0.25	0.910813	0.862786	0.437949	0.920376	0.923658	0.47937	
4	-0.002918	0.833333	0.921397	0.75	0.910813	0.862786	0.437949	0.920376	0.923658	0.47937	

Figure 8 "Feature Scaling applied on categorical features"

3.2.6 Feature Selection

Feature selection is applied on the OWID dataset. The top motivations to utilize feature selection are that it empowers the ML model to execute quicker. It lessens the intricacy of a model and makes it simpler to decipher. Moreover ,it improves the accuracy when the right features are picked. It can be shown in figure. 9.



```
In [32]: # let's print the number of total and selected features

         # this is how we can make a list of the selected features
         selected_feat = X_train.columns[(feature_sel_model.get_support())]

         # let's print some stats
         print('total features: {}'.format((X_train.shape[1])))
         print('selected features: {}'.format(len(selected_feat)))

         total features: 113
         selected features: 17

In [33]: selected_feat

Out[33]: Index(['location', 'total_deaths', 'total_cases_per_million',
                'stringency_index', 'population', 'total_casesnan', 'total_deathsnan',
                'total_deaths_per_millionnan', 'reproduction_ratenan',
                'new_tests_smoothednan', 'stringency_indexnan', 'extreme_povertynan',
                'cardiovasc_death_ratenan', 'diabetes_prevalencenan',
                'hospital_beds_per_thousandnan', 'human_development_indexnan',
                'date_encoded'],
               dtype='object')
```

Figure 9 "Feature Selection on the data"

3.2.7 Application of Regression Model

Regression analysis is a dependable strategy for recognizing which factors affect a subject of interest. The method involves splitting the data into 20% test and 80% train sets. We train the model using the training data so it can later on be tested for its generalization using the testing data. A multiple linear regression model has been applied to the OWID covid-19 data set to predict the next wave of coronavirus disease. A multiple linear regression model is a mathematical and statistical technique that uses several explanatory variables to predict the outcome of a single variable [44]. Formula for the calculation of multiple linear regression model is given as follow:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

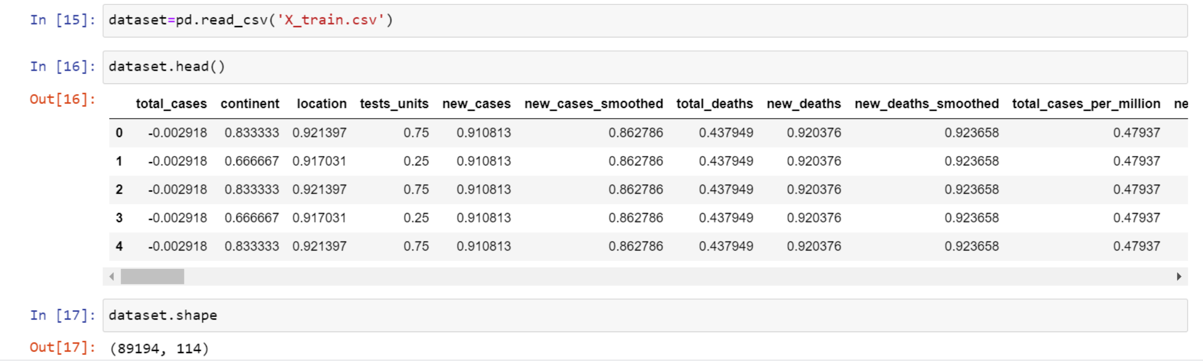
x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

The model is given 80% of data for training purpose:

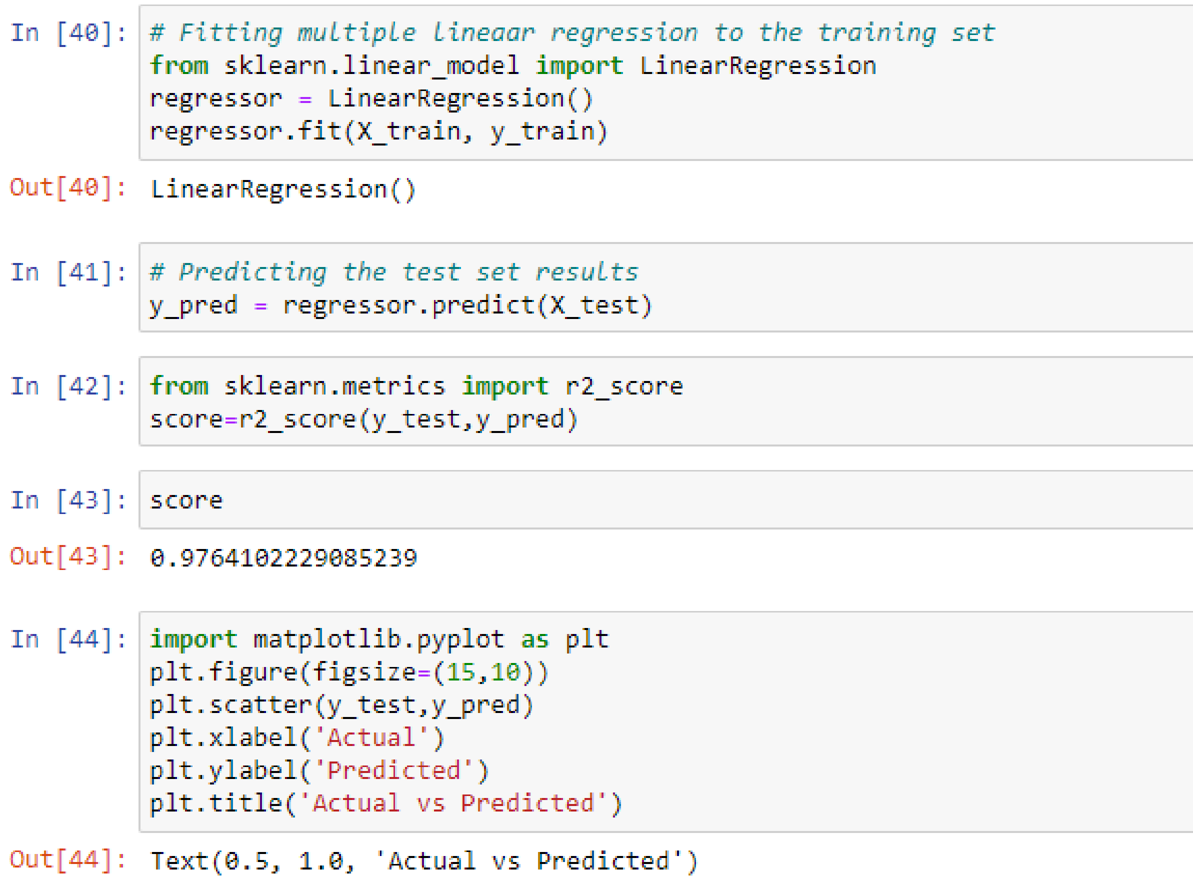


```
In [15]: dataset=pd.read_csv('X_train.csv')

In [16]: dataset.head()

Out[16]:
```

	total_cases	continent	location	tests_units	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	ne
0	-0.002918	0.833333	0.921397	0.75	0.910813	0.862786	0.437949	0.920376	0.923658	0.47937	
1	-0.002918	0.666667	0.917031	0.25	0.910813	0.862786	0.437949	0.920376	0.923658	0.47937	
2	-0.002918	0.833333	0.921397	0.75	0.910813	0.862786	0.437949	0.920376	0.923658	0.47937	
3	-0.002918	0.666667	0.917031	0.25	0.910813	0.862786	0.437949	0.920376	0.923658	0.47937	
4	-0.002918	0.833333	0.921397	0.75	0.910813	0.862786	0.437949	0.920376	0.923658	0.47937	

```
In [17]: dataset.shape

Out[17]: (89194, 114)
```

The application of the model is given below in figure 10 and 10.1.



```
In [40]: # Fitting multiple lineaar regression to the training set
         from sklearn.linear_model import LinearRegression
         regressor = LinearRegression()
         regressor.fit(X_train, y_train)

Out[40]: LinearRegression()

In [41]: # Predicting the test set results
         y_pred = regressor.predict(X_test)

In [42]: from sklearn.metrics import r2_score
         score=r2_score(y_test,y_pred)

In [43]: score

Out[43]: 0.9764102229085239

In [44]: import matplotlib.pyplot as plt
         plt.figure(figsize=(15,10))
         plt.scatter(y_test,y_pred)
         plt.xlabel('Actual')
         plt.ylabel('Predicted')
         plt.title('Actual vs Predicted')

Out[44]: Text(0.5, 1.0, 'Actual vs Predicted')
```

Figure 10 “Regression Model”

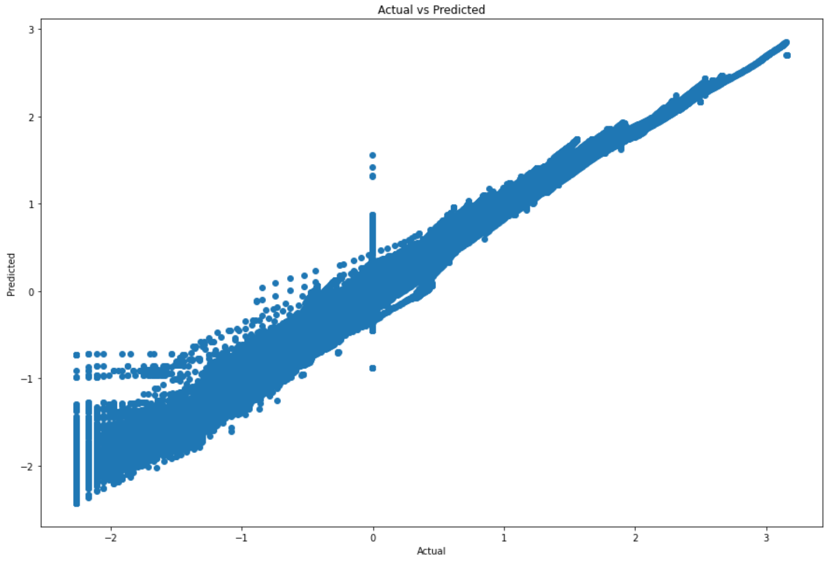


Figure 10.1 – Actual vs Predicted

Figure 10.1 shows represents results achieved after employing Multiple Linear Regression on test data which was left as raw data after segregation. The graph shows promising results and with passage of time and data the results are achieving an optimal level.

Chapter 4

Results & Discussions

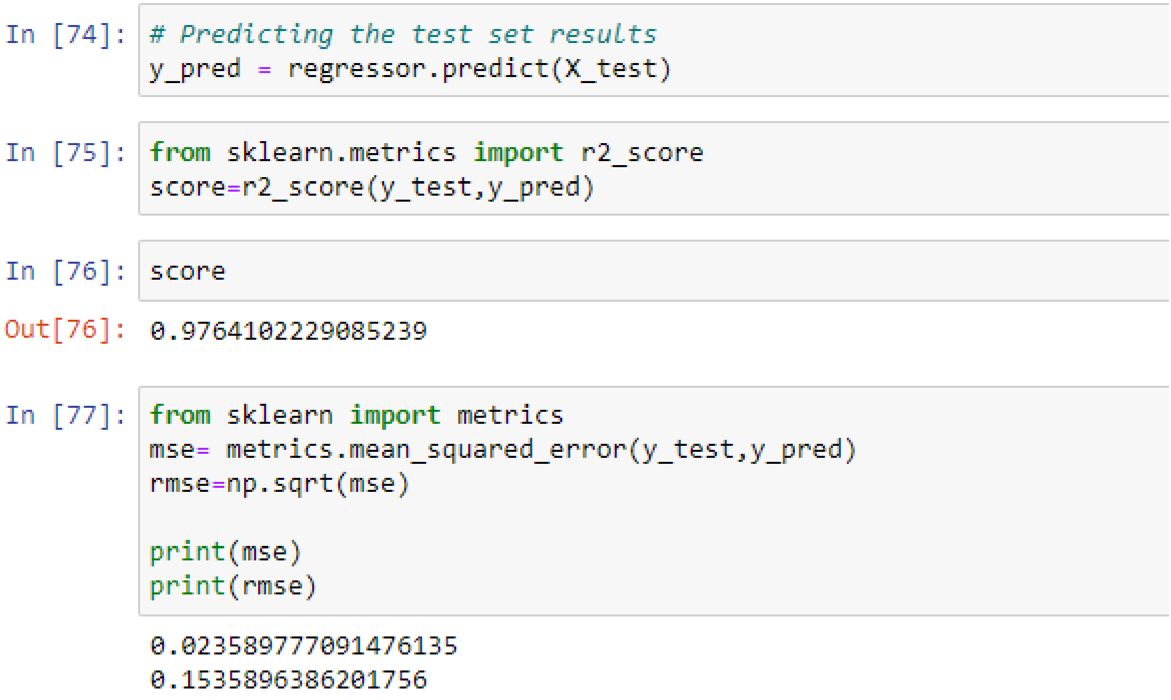
CHAPTER 4: RESULTS & DISCUSSION

After applying the regression model on the OWID covid-19 dataset , the following results have been obtained as shown in the tables below. The outcome of the model determines that the model fitted well on the OWID covid-19 data set. The perspective of testing dataset of the model has an accuracy of 0.97 i.e 97.6%. Moreover, the mean square error (MSE) and root mean square error (RMSE) scores of the regression model are 0.0235 and 1.5358, respectively.

Data	MSE	RMSE	R2 Score	% Accuracy
Total_cases	0.0235	1.5358	0.97641	97.6%

Table 2 "Results of the regression model"

Both MSE and RMSE are used as the evaluating metrics of the regression model's performance i.e to check its reliability in predicting the outcome.



```
In [74]: # Predicting the test set results
         y_pred = regressor.predict(X_test)

In [75]: from sklearn.metrics import r2_score
         score=r2_score(y_test,y_pred)

In [76]: score

Out[76]: 0.9764102229085239

In [77]: from sklearn import metrics
         mse= metrics.mean_squared_error(y_test,y_pred)
         rmse=np.sqrt(mse)

         print(mse)
         print(rmse)

         0.023589777091476135
         0.1535896386201756
```

Figure 11 "Results of the regression model"

Statistical models such as Regression models are significant methods for assessing transmittable virus data analyses at a specific point in time. During the literature review it was highlighted that most of the reviewed work has applied models directly to the data collected and aren't generalized i.e they predict the wave based on a specific country. It was observed that most of the work that has been done doesn’t handle missing values present in the data. Missing values can jeopardize the efficiency of the model and the accuracy of the results. Moreover, since most of the data which is collected is skewed. If the skewed data isn’t transformed it will fail to satisfy the homogeneity of the variance for the error and the model will fail to fit the linear. It can be deduced easily while looking at table 3 that most of the models did not apply data transformation. In addition, the majority of the models fail to perform feature processing which is a crucial step in determining which features will give an optimal solution when chosen. Apart from that only one other model has applied feature scaling to normalize the range of data. Furthermore, no model has applied feature selection as it is an important step towards model fitting. Lastly, most of the models aren’t generalized i.e they do not predict the wave based on the data from all the countries but a specific country. This clearly culminates that there are a lot of gaps present in the work of others and our model tries to fulfill these gaps to predict the future with an accuracy of 97.6%.

Gaps	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	Our model
Handling missing values	no	yes	no	no	no	no	no	no	no	no	yes
Data Transformation	no	no	no	no	no	no	no	no	no	no	yes
Feature Extraction	no	no	yes	no	no	no	no	no	no	no	no
Feature preprocessing	no	yes	yes	no	no	no	no	no	no	no	yes
Feature Scaling	yes	no	no	no	no	no	no	no	no	no	yes
Feature Selection	no	no	no	no	no	no	no	no	no	no	yes
General model	no	no	no	yes	no	yes	no	no	yes	no	yes

Table 3 "Comparison of models"

Chapter 5

Model Evaluation

CHAPTER 5: MODEL EVALUATION

Regression model is a predictive modelling technique which can be used to probe the correspondence between independent and dependent variables. This technique can be used in predictions, forecasting, time series modelling and to find the cause and effect relationship between variables [45]. Model evaluation is vital in prediction models. It assists you with understanding the presentation of your model and makes it simple to introduce your model to others. There are a wide range of assessment measurements out there yet just some of them are reasonable to be utilized for relapse.

There are 3 principle measurements for model assessment in relapse:

1. R Square/Adjusted R Square

2. Mean Square Error(MSE)

3. Root Mean Square Error(RMSE)

For the evaluation of this regression model we have utilized all of the above mentioned metrics.

5.1 R square Evaluation:

R Square estimates how much inconsistency in subordinate variables can be clarified by the model. It is the square of the Correlation Coefficient(R) and that is the reason it is called R Square.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

R Square is determined by the amount of squared of expectation mistake separated by the all out amount of the square which replaces the determined forecast with mean. R Square worth is between 0 to 1 and a greater worth demonstrates a superior fit among expectation and genuine worth.

5.2 MSE Evaluation:

Mean Square Error is an outright proportion of the decency for the fit.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

MSE is determined by the amount of square of expectation blunder which is genuine yield less anticipated yield and afterward partition by the quantity of information focused. It gives you an outright number on how much your anticipated outcomes stray from the genuine number.

5.3 RMSE Evaluation:

Root Mean Square Error(RMSE) is the square foundation of MSE. It is utilized more generally than MSE in light of the fact that initially some of the time MSE worth can be too enormous to even think about contrasting without any problem. Furthermore, MSE is determined by the square of blunder, and in this manner square root takes it back to a similar degree of forecast mistake and makes it simpler for translation.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

Chapter 6

Conclusion & Future Work

CHAPTER 6: CONCLUSION & FUTURE WORK

6.1 Conclusion

In this study, systematic efforts are made to design a methodology that results in the prediction of the next wave of the deadly coronavirus. Amid this work, a machine learning regression algorithm was employed i.e. Multiple Linear Regression which uses multiple independent variables to predict the outcome of a single dependent variable and was evaluated based on R square. MSE and RMSE. Prediction was derived based on the OWID covid-19 dataset with an accuracy of 97.6% along with 0.0235 MSE, 1.5358 RMSE and 0.97641 R^2 based on total cases. It can be concluded that the model will be able to predict the future accurately with an accuracy of 97.6%. During the literature review it was discerned that no work has been done regarding the feature selection and engineering in any of the reviewed papers, moreover, the accuracy of models for prediction reviewed had maximum accuracy of 97%. While using machine learning algorithms the preprocessing of data is a very crucial step to achieve accurate and optimal outcomes. The data was separated into categorical and numerical features to tackle their missing values accordingly. Another gap that was identified was that none of the work reviewed was handling the skewness in the data. Our model applied different transformation techniques to highlight the skewness present in the data and uses yoe johnson's technique to normally distribute the data. Moreover, min-max feature scaling is being implemented in our model to scale the features on a specific range which is yet another gap identified in the work reviewed. Our model also makes use of feature selection which is one of the ambiguities highlighted in the reviewed papers as feature selection helps in the faster execution of the model. Out of 113 attributes, 17 attributes are used to build a regression model for the prediction of the virus after which we performed model fitting to achieve optimal results. Furthermore, most of the reviewed work makes predictions based on specific countries whereas the proposed model is making generalized predictions taking various continents and locations into consideration. In conclusion, overall our model has a more generalized approach towards the prediction of occurrence of the next wave.

6.2 Future Work

In future we can incorporate support vector regression (SVR) to achieve maximum accuracy and for the model to perform well. Another future dimension is utilizing more feature selection techniques for our model to perform better. Furthermore, we can preprocess the data differently to mitigate missing values or noise in the data. Moreover, other evaluation techniques can be used. Lastly we can use more data as in ML the more data you have the better results you get. In the future, the proposed methodology with the utilized machine learning algorithms can help governments to proactively make necessary arrangements, identify hotspots for disaster management and to help prioritize and plan national activities for law enforcement institutions, along with international organizations to proactively make decisions to contain the spread by restricting travels in such regions.

References

REFERENCES

1. Elaziz, M. A., Hosny, K. M., Salah, A., Darwish, M. M., Lu, S., & Sahlol, A. T. (2020). New machine learning method for image-based diagnosis of COVID-19. Plos one, 15(6), e0235187.

2. He, X., Yang, X., Zhang, S., Zhao, J., Zhang, Y., Xing, E., & Xie, P. (2020). Sample-Efficient Deep Learning for COVID-19 Diagnosis Based on CT Scans.

3. Sufian, A., Ghosh, A., Sadiq, A. S., & Smarandache, F. (2020). A survey on deep transfer learning to edge computing for mitigating the covid-19 pandemic. Journal of Systems Architecture, 108, 101830.

4. Rahman, M. A., Zaman, N., Asyhari, A. T., Al-Turjman, F., Bhuiyan, M. Z. A., & Zolkipli, M. F. (2020). Data-driven dynamic clustering framework for mitigating the adverse economic impact of Covid-19 lockdown practices. Sustainable cities and society, 62, 102372.

5. Ahmad, Amir, et al. "The number of confirmed cases of covid-19 by using machine learning: Methods and challenges." Archives of Computational Methods in Engineering (2020): 1-9.

6. University of Illinois at Chicago. (2021, July 22). Machine Learning in Healthcare: Examples, Tips & Resources for Implementing into Your Care Practice. UIC Online Health Informatics.

7. Coronavirus Disease (COVID-19): What Is It, Symptoms, Causes & Prevention. (n.d.). Cleveland Clinic. Retrieved September 10, 2021, from https://my.clevelandclinic.org/health/diseases/21214-coronavirus-covid-19

8. Human Coronavirus Types | CDC. (n.d.). National Center for Immunization and Respiratory Diseases (NCIRD). Retrieved September 10, 2021, from https://www.cdc.gov/coronavirus/types.html

9. Parbat, D., & Chakraborty, M. (2020). A python-based support vector regression model for prediction of COVID19 cases in India. Chaos, Solitons & Fractals, 138, 109942.

10. Rath, S., Tripathy, A., & Tripathy, A. R. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression models. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(5), 1467-1474.

11. Gothai, E., Thamilselvan, R., Rajalaxmi, R. R., Sadana, R. M., Ragavi, A., & Sakthivel, R. (2021). Prediction of covid-19 growth and trend using machine learning approach. Materials Today: Proceedings.

12. Singhal, A., Singh, P., Lall, B., & Joshi, S. D. (2020). Modeling and prediction of COVID-19 pandemic using Gaussian mixture model. Chaos, Solitons & Fractals, 138, 110023.

13. Tomar, A., & Gupta, N. (2020). Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Science of The Total Environment, 728, 138762.

14. Maleki, M., Mahmoudi, M. R., Wraith, D., & Pho, K. H. (2020). Time series modeling to forecast the confirmed and recovered cases of COVID-19. Travel medicine and infectious disease, 37, 101742.

15. Tandon, H., Ranjan, P., Chakraborty, T., & Suhag, V. (2020). Coronavirus (COVID-19): ARIMA-based time-series analysis to forecast the near future. arXiv preprint arXiv:2004.07859.

16. Arora, P., Kumar, H., & Panigrahi, B. K. (2020). Prediction and analysis of COVID-

19 positive cases using deep learning models: A descriptive case study of India. Chaos, Solitons & Fractals, 139, 110017.

17. Alazab, M., Awajan, A., Mesleh, A., Abraham, A., Jatana, V., & Alhyari, S. (2020). COVID-19 prediction and detection using deep learning. International Journal of Computer Information Systems and Industrial Management Applications, 12, 168-181.

18. Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., & dos Santos Coelho, L. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos, Solitons & Fractals, 135, 109853.

19. COVID Live Update: 224,130,821 Cases and 4,622,908 Deaths from the Coronavirus - Worldometer. (n.d.). Worldometer. Retrieved September 10, 2021, from https://www.worldometers.info/coronavirus/

20. Zhang X, Ma R, Wang L. Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries. Chaos Solitons Fractals. 2020;135:109829. doi:10.1016/j.chaos.2020.109829

21. Lamiaa, AmarAshraf A. TahaMarwa Y. Mohamed (June 2020) . Prediction of thefinal size for COVID-19 epidemic usingmachine learning: A case study of Egypt

22. Education, I. C. (2021, July 6). Exploratory Data Analysis. Ibm. https://www.ibm.com/cloud/learn/exploratory-data-analysis

23. Kaushik, S. (2020, October 19). Feature Selection Methods | Machine Learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/#:%7E:text=Top%20reasons%20to%20use%20feature,the%20right%20subset%20is%20chosen.

24. Wynants L, Van Calster B, Collins G S, Riley R D, Heinze G, Schuit E et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal BMJ 2020; 369 :m1328 doi:10.1136/bmj.m1328

25. Kamal, M., Aljohani, A., & Alanazi, E. (2020). IoT meets COVID-19: Status, Challenges, and Opportunities. arXiv preprint arXiv:2007.12268. https://arxiv.org/pdf/2007.12268v1.pdf

26. Yadav, M., Perumal, M., & Srinivas, M. (2020). Analysis on novel coronavirus (COVID-19) using machine learning methods. Chaos, Solitons & Fractals, 139, 110050. https://www.sciencedirect.com/science/article/pii/S0960077920304471?casa_token=jkIp8GORe48AAAAA:PrOeCj1_6icvEcx8fJidplZEkRuctSvvOq5H566eCCFWWUNwHiw9T1jU_0xoYvedQEt93i-u5jk

27. Kompella, V., Capobianco, R., Jong, S., Browne, J., Fox, S., Meyers, L., ... & Stone, P. (2020). Reinforcement Learning for Optimization of COVID-19 Mitigation policies. arXiv preprint arXiv:2010.10560.

28. Shoeibi, A., Khodatars, M., Alizadehsani, R., Ghassemi, N., Jafari, M., Moridian, P., ... & Alizadehsani, Z. (2020). Automated detection and forecasting of covid-19 using deep learning techniques: A review. arXiv preprint arXiv:2007.10785.

https://arxiv.org/pdf/2007.10785

29. Bashir, A., Izhar, U., & Jones, C. (2020). IoT-Based COVID-19 SOP Compliance and Monitoring System for Businesses and Public Offices. In Engineering Proceedings (Vol. 2, No. 1, p. 14). Multidisciplinary Digital Publishing Institute. https://www.mdpi.com/2673-4591/2/1/14

30. Abir, S. M., Islam, S. N., Anwar, A., Mahmood, A. N., & Oo, A. M. T. (2020). Building Resilience against COVID-19 Pandemic Using Artificial Intelligence, Machine Learning, and IoT: A Survey of Recent Progress. IoT, 1(2), 506-528. https://www.mdpi.com/2624-831X/1/2/28

31. Bian, Sizhen, et al. "A wearable magnetic field based proximity sensing system for monitoring COVID-19 social distancing." Proceedings of the 2020 International Symposium on Wearable Computers. 2020.

https://dl.acm.org/doi/abs/10.1145/3410531.3414313

32. Kumar, K., Kumar, N., & Shah, R. (2020). Role of IoT to avoid spreading of COVID-19. International Journal of Intelligent Networks, 1, 32-35. https://www.sciencedirect.com/science/article/pii/S2666603020300026

33. Petrović, N., & Kocić, Đ. (2020). IoT-based System for COVID-19 Indoor Safety Monitoring. preprint), IcETRAN, 2020, 1-6.

https://www.researchgate.net/profile/Nenad-Petrovic/publication/343231422_IoT-based_System_for_COVID-19_Indoor_Safety_Monitoring/links/5f67315b458515b7cf418f2b/IoT-based-System-for-COVID-19-Indoor-Safety-Monitoring.pdf

34. Vaid, Shashank, Caglar Cakan, and Mohit Bhandari. "Using machine learning to estimate unobserved COVID-19 infections in North America." The Journal of bone and joint surgery. American volume (2020).

https://link.springer.com/article/10.1007/s11831-020-09472-8

35. Adlen Ksentini, Bouziane Brik. An Edge-Based Social Distancing Detection Service to Mitigate COVID-19 Propagation. IEEE Internet of Things Magazine (2020).

https://ieeexplore.ieee.org/abstract/document/9241469/

36. Swapnili Karmore, Rushikesh Bodhe, Fadi Al-Turjman, R Lakshmana Kumar, Sofia Pillai. IoT Based Humanoid Software for Identification and Diagnosis of Covid-19 Suspects. IEEE Sensors Journal (2020).

https://ieeexplore.ieee.org/abstract/document/9234596

37. Mohammed, M. N., Syamsudin, H., Al-Zubaidi, S., AKS, R. R., & Yusuf, E. (2020). Novel COVID-19 detection and diagnosis system using IOT based smart helmet. International Journal of Psychosocial Rehabilitation, 24(7).

https://www.researchgate.net/profile/Assoc_Prof_Dr_Mohammed_Abdulrazaq/publication/340264439_NOVEL_COVID-19_DETECTION_AND_DIAGNOSIS_SYSTEM_USING_IOT_BASED_SMART_HELMET/links/5e8dba3ba6fdcca789fe053e/NOVEL-COVID-19-DETECTION-

AND-DIAGNOSIS-SYSTEM-USING-IOT-BASED-SMART-HELMET.pdf

38. Ahammed, K., Satu, M. S., Abedin, M. Z., Rahaman, M. A., & Islam, S. M. S. (2020). Early Detection of Coronavirus Cases Using Chest X-ray Images Employing Machine Learning and Deep Learning Approaches. medRxiv.

https://www.medrxiv.org/content/10.1101/2020.06.07.20124594v1

39. Rashid, M. T., & Wang, D. (2020). CovidSens: a vision on reliable social sensing for COVID-19. Artificial Intelligence Review, 1-25.

https://arxiv.org/abs/2004.04565

40. Batra, R., Chan, H., Kamath, G., Ramprasad, R., Cherukara, M. J., & Sankaranarayanan, S. K. (2020). Screening of therapeutic agents for COVID-19 using machine learning and ensemble docking studies. The journal of physical chemistry letters, 11(17), 7058-7065.

https://pubs.acs.org/doi/abs/10.1021/acs.jpclett.0c02278

41. Khalifa, N. E. M., Smarandache, F., Manogaran, G., & Loey, M. (2020). A study of the neutrosophic set significance on deep transfer learning models: An experimental case on a limited covid-19 chest x-ray dataset. Cognitive Computation, 1-10. https://link.springer.com/article/10.1007/s12559-020-09802-9

42. H. Yin, B. Mukadam, X. Dai, and N. Jha, "DiabDeep: Pervasive Diabetes Diagnosis based on Wearable Medical Sensors and Efficient Neural Networks," IEEE Transactions on Emerging Topics in Computing, 2019. https://www.researchgate.net/publication/341980921

43. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available:

https://doi.org/10.1109/CVPR.2016.90

44. Kukar, M., Gunčar, G., Vovko, T., Podnar, S., Černelč, P., Brvar, M., ... & Notar, M.

(2020). COVID-19 diagnosis by routine blood tests using machine learning. arXiv preprint arXiv:2006.03476.

https://arxiv.org/ftp/arxiv/papers/2006/2006.03476.pdf

45. Magar, R., Yadav, P., & Farimani, A. B. (2020). Potential neutralizing antibodies discovered for novel coronavirus using machine learning. arXiv preprint arXiv:2003.08447.

https://arxiv.org/ftp/arxiv/papers/2003/2003.08447.pdf

46. Sethy, P. K., Behera, S. K., Ratha, P. K., & Biswas, P. (2020). Detection of coronavirus disease (COVID-19) based on deep features and support vector machines.

https://www.preprints.org/manuscript/202003.0300/v2

47. M. Putra, Z. Yussof, K. Lim, and S. Salim, “Convolutional neural network for person and car detection using yolo framework,” Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 10, no. 1-7, pp. 67–71, 2018.

https://arxiv.org/pdf/2005.01385.pdf

48. Barstugan, M., Ozkaya, U., & Ozturk, S. (2020). Coronavirus (covid-19) classification using ct images by machine learning methods. arXiv preprint arXiv:2003.09424.

49. Zhaozhi Qian, Ahmed M. Alaa, Mihaela van der Schaar. (2020). CPAS: the UK's National Machine Learning-based Hospital Capacity Planning System for COVID-19. arXiv preprint arXiv:2007.13825.

50. Oluwasegun A. Somefun, Folasade Dahunsi. (2020). From the logistic-sigmoid to nlogistic-sigmoid: modelling the COVID-19 pandemic growth. https://arxiv.org/pdf/2008.04210v3.pdf

51. Mauricio Arango. (2020). COVID-19 Pandemic Cyclic Lockdown Optimization Using Reinforcement Learning. https://arxiv.org/ftp/arxiv/papers/2009/2009.04647.pdf

52. Giuseppe Carlo Calafiore, Carlo Novara, Corrado Possieri. (2020). A Modified SIR Model for the COVID-19 Contagion in Italy. https://arxiv.org/pdf/2003.14391.pdf

53. Barstugan, M., Ozkaya, U., & Ozturk, S. (2020). Coronavirus (covid-19) classification using ct images by machine learning methods. arXiv preprint arXiv:2003.09424.https://arxiv.org/abs/2003.09424

54. Danda, Shashank Reddy, and Bo Chen. "Toward Mitigating Spreading of Coronavirus via Mobile Devices." IEEE Internet of Things Magazine 3.3 (2020): 12-16.

55. Liu, Dianbo, et al. "A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models." arXiv preprint arXiv:2004.04019 (2020).

56. Cheng, Fu-Yuan, et al. "Using machine learning to predict ICU transfer in hospitalized COVID-19 patients." Journal of clinical medicine 9.6 (2020): 1668.

57. *Multiple Linear Regression (MLR) Definition*. (n.d.). Investopedia. Retrieved September 10, 2021, from https://www.investopedia.com/terms/m/mlr.asp#:%7E:text=Key%20Takeaways-,Multiple%20linear%20regression%20(MLR)%2C%20also%20known%20simply%20as%20multiple,uses%20just%20one%20explanatory%20variable.

58. Ray, S. (2020, October 18). *Regression Techniques in Machine Learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/#:%7E:text=Regression%20analysis%20is%20a%20form,effect%20relationship%20between%20the%20variables.

59. Wu, S. (2021, June 5). *3 Best metrics to evaluate Regression Model? - Towards Data Science*. Medium. https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b

60. Dogan, O. (2021b, July 5). *A systematic review on AI/ML approaches against. . .* Complex & Intelligent Systems. https://idp.springer.com/authorize?response_type=cookie&client_id=springerlink&redirect_uri=https%3A%2F%2Flink.springer.com%2Farticle%2F10.1007%2Fs40747-021-00424-8

61. Wang, L., Li, J., Guo, S., Xie, N., Yao, L., Cao, Y., ... & Sun, D. (2020). Real-time estimation and prediction of mortality caused by COVID-19 with a patient information based algorithm. *Science of the total environment*, *727*, 138394.

62. Tandon, H., Ranjan, P., Chakraborty, T., & Suhag, V. (2020). Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. *arXiv preprint arXiv:2004.07859*.

63. Fu, L., Wang, B., Yuan, T., Chen, X., Ao, Y., Fitzpatrick, T., ... & Zou, H. (2020). Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: a systematic review and meta-analysis. *Journal of Infection*, *80*(6), 656-665.