National College of Ireland

Forecasting and Analysis of COVID-19 Pandemic

MSc Research Project
Data Analytics

Kanak Kaushik
Student ID: 18136966

School of Computing
National College of Ireland

Supervisor: Mr. Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Kanak Kaushik
Student ID:	X18136966
Program:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Mr. Jorge Basilio
Submission Due Date:	23/4/2020
Project Title:	Forecasting and Analysis of COVID-19 Pandemic
Word Count:	5743
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to the research I conducted for this project. All information other than my contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	23rd April 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not enough to keep a copy on the computer.	

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Forecasting and Analysis of COVID-19 Pandemic

Kanak Kaushik
X18136966

Abstract

COVID-19 is also known as Novel Coronavirus, was first found at a wet market in Wuhan, China. As the upsurge of the COVID-19 affects the respiratory system caused by severe acute respiratory syndrome (SARS) virus advances within the world, an increase in epidemiological data ensures researchers to make a plan of action for societal awareness and precautions against the virus. Machine learning algorithms can be applied to available datasets to explore insights that will help the countries to be prepared. This research uses multiple machine learning algorithms to predict the infected cases of coronavirus all over the world and analysis of the datasets. Machine learning algorithms like Support Vector Regressor have the lowest R^2 score = 0.8273 among Polynomial Regression, and Bayesian Ridge Regression, and the highest RMSE value = 124328.5297 amongst the three models, which tells us that the Support Vector Regressor is the least preferred model to choose. Bayesian Ridge Regression has R^2 score = 0.9321 and the lowest RMSE value = 71920.7332 to be the best model among the three.

Keywords: *COVID-19, SARS, Machine learning, data analysis, SVR, Regression*

1. Introduction

In present-day humankind's history, we have seen multiple infectious disease outbreak, which posed a threat to public health repeatedly. Recently a novel coronavirus was discovered in Wuhan, China, which attacks the respiratory system and has severe acute respiratory syndrome (SARS). It is a highly infectious disease that can transmit from human to human and even can live onto surfaces. It is named as COVID-19, in which 19 denotes the year 2019 in which it was discovered (Zhang *et al.*, 2020). The outbreak started in china slowly seeping into the world, causing the world to panic. Studies revealed that all ages are vulnerable to COVID-19 infections, even some cases falling to fatal respiratory diseases.

Similar to this virus, there has been similar coronavirus outbreak in the recent era, like SARS and Middle East Respiratory Syndrome (MERS) in 2003 and September 2012, respectively. SARS infected around 8000 people in its lifetime, and 10 percent of the infected people died from it. Still, there is no cure or vaccine developed against the virus. The virus was contained in 2004, and no cases have appeared until now. MERS is also known as camel-flu, was first discovered in 2012 in Middle East countries (‘Middle East respiratory syndrome’, 2020). It has an incubation period of 2-14 days from the first exposure of the virus. It has infected around 2500 people worldwide and has taken 866 people to death. Both of these diseases are infectious means it can spread very quickly through the human transmission. Like SARS, no vaccines are developed against the disease.

COVID-19 is also a strain of coronaviruses means it has a similar family-like SARS and MERS. It is a novel coronavirus as there has not been any virus like this before. Its symptoms are fever, cough, and shortness of breath. The first registered case was in Wuhan, and then it slowly spread across the world (Sedaghat *et al.*, 2020). The incubation period is like MERS of 2-14 days and has three classifications of the symptoms like mild, severe, and fatal. World Health Organization classified the outbreak as pandemic on 11th March 2020. To prevent the spread of the virus more, most of the countries have adapted social distancing, and instructed people to stay-at-home and go outside only if needed. Around 2.6 million people have been infected, resulting in approximately 179,000 deaths, which will increase as the days are passed. Every country is trying to "Flatten the Curve", which means decreasing the infection rate to cope up with the hospital administration without being overwhelmed, which can cause unnecessary fatalities. A similar pandemic outbreak was of 1918 Spanish flu, caused havoc on the world system, and had a brutal impact on the systems. Vaccines are currently in the works and will likely be available in the last quarter of 2020.

All pandemics have a massive impact on the society, politics, and economy of the country. Epidemics can affect temporary as well as permanent damage to the socio-economic growth of the country. Layoffs, temporary closure of shops, agriculture, manufacturing sectors, services, and many more disruptions are affecting the economic difficulties for private, public, and government institutions (Madhav *et al.*, 2017). Social impacts like the restriction of civil movements for travel, closing of schools, recreational centers closing, and more effects on the society (Qiu *et al.*, 2017).

The essential factor in the pandemic is the reproduction rate (R0) of the virus. It shows the transmission of the virus into the society. It is vital to keep the R0 rate as low as possible to contain the outbreak of the virus. If (R0=5), then every person will infect five people who will affect another five people, resulting in the exponential growth of the virus (Ding *et al.*, 2020).

The objective of this research is to attempt to predict the number of COVID-19 cases all around the world and analysis of the datasets to gain some information on the current situation.

1.1 Research Question

This study tries to find the answer to the following research question:
"*Can machine learning algorithms help in accurately predicting the COVID-19 cases in the world?*"

This research report consists of a review of previous work done in the field, the methodology of the research, implementation, evaluation, and conclusion of the report.

2. Related Work

2.1 Introduction

Data/information is growing exponentially every day, making it harder to comprehend and look for specific information. This bundle of data can be used to gain some insights by analyzing the data and recommending the knowledge gain by the analysis to the business or marketplace to improve the strategies and give the most out of it to the people. The conversion of raw information into some knowledgeable content that makes acute observations from the statistics is known as data analysis. Statistical analytics helps in like

recommendation, forecasting, spam detection, and many more. These analyses can show insights into the data, which would be near impossible to get from the datasets. Some statistics analytics are Diagnostic analytics, Prescriptive analytics, Predictive Analytics, and Descriptive analytics (Frankenfield, 2019). This research will be based on forecasting the values inferred from the past values known as predictive analysis. Businesses use a lot of prediction to gain an advantage in sales and serve the customer. From this knowledge, the company takes an edge over its competitors who are unaware of this technology (Department of Information Technology et al., 2019).

2.2 Literature Review

This part acknowledges in the field of pandemic predictions and its analysis. Researches have done a study on this topic and have presented their findings.

2.2.1 Prediction models for Epidemics:

(Iannelli and Pugliese, 2014) used different mathematical models to see the patterns in the events of the epidemic. They described three classes of infections: susceptible to infections (S), infected individuals who can transmit the virus onto others (I), and the immune individual who already had been infected (R). They developed the model based on these characters' SIR models and even explained the concept of the reproductive rate of the infectious virus. They concluded that if the virus falls below the survival reproduction rate, then $R_0 = 1$, which means the infected person only transmits it to the other one person. And even herd immunity can be achieved if the virus has infected a substantial amount of the people, and the rate of vaccination must be higher to attain the fairest result in the epidemic.

Another study done by (Raissi, Ramezani and Seshaiyer, 2019), they argued that even though the SIR model is efficient and has been used for a very long time, they thought introducing different model parameters can result into a better throughput result. They introduced the Nelder-Meade based heuristic method in the modeling. They found there are two main contributors in modeling the infectious diseases, deep learning and statistical inferences based on time-series model. They started by introducing these techniques; the results would be more accurate and less computational power. They proved that the proposed approaches are better for parameter estimation and prediction of the expanse of the virus.

In another study, the researchers focused on predicting R_0, reproduction rate of the disease, a factor influencing the disease spreadability. They used three types of machine learning models to determine which fit the best and accurate. They trained the model on test data and predicted R_0 values of the test data. Support vector regressor and Artificial Neural Network showed better accuracy and fit of the model rather than the linear model. They concluded that the spread of the disease could be predicted using global properties, and standard machine learning algorithms are effective on the datasets (Tripathi, Reza and Garg, 2019).

Ensemble methods were used by (Ray and Reich, 2018) to get better and accurate results from the previous works. Ensemble methods mean the amalgamation of multiple models, and passing the variables through it to get a single prediction is known as ensemble methods. It has the strength of every mixed model, therefore giving more confidence in the predictions. In this study, they computed weights average on factors like peak intensity of the virus, seasonal timing of the virus, affecting people’s health. There were two ensemble approach CW and FW-reg-w models. FW-reg-w ensemble outperformed CW ensemble in

good average and worst-case performance, while CW performed similarly, but its worst-case performance was lacking than the FW-reg-w ensemble model.

Another ensemble method approach was studied by Chowell *et al.*, 2020; they introduced an ensemble model for sequential forecasting for the Ebola Forecasting challenge. Generalized Growth Model (GGM) and Generalized Logistic Model (GGL) were used in this study. They tested their model based on root-mean-squared-error (RMSE) to quantify between the models. GLM outperformed all the other models in forecasting 1 – 4 weeks of the period. Even the GGM-GGL ensemble outperformed many models by comparing the RMSE values. Their ensemble models outperformed every other participant model in the challenge.

Therefore, many studies have used different models for the study of epidemics.

2.2.2 Prediction models for SARS and MERS:

2.2.2.1 SARS

The study was done to find the important features of the SARS epidemic in Beijing 2003 using space-time statistics. (Wang *et al.*, 2008) found that spatial neighborhood is an important factor of the epidemic spread model, although population density transmission and healthcare workers were occasionally significant. Their study showed that enhanced control measures at the minute area are the most effective approach in the outbreak. Measures like quarantine, social distancing were important factors in the decline of the epidemic. This study added to include spatial variation in data to increase the efficient involvement of the measures taken into any epidemic scenarios.

2.2.2.2 MERS

Middle East Respiratory Syndrome (MERS) is a strain of coronaviruses that had an outbreak in the year 2013. It attacks the respiratory system leaving them vulnerable or even fatal in some cases. Its symptoms are mild to moderate cold. Data mining techniques are applied to the datasets to understand the recovery from the MERS. (Al-Turaiki, Alshahrani and Almutairi, 2016) did study on building predictive models for MERS. They implemented by using two machine learning algorithms like Naïve Bayes classifier and J48 decision tree to use against the datasets to analyze the data. As this study was based on classification, they used evaluation measures like accuracy, precision, and recall evaluating the models. Their study concluded that the Naïve Bayes recovery model outperforms with better overall accuracy than the J48 decision tree algorithm. But J48 performs better instability model than Naïve Bayes.

Another study was done by (John and Shaiba, 2019) used SVM, Naïve Bayes, CTree, and J48 algorithms to classify the factors that were influential in recovering the MERS infected individuals. J48 decision tree algorithm put a light on the recovery of the MERS patient was heavily influenced by if there were any underlying health conditions and if the patient healthcare professional or not. They concluded that attributes like underlying health conditions, age, early condition of the infected person, and if a healthcare professional or not have a huge impact on the recovery rate of the infected person. Thus, machine learning algorithms helped the researchers to shed some light on the various attributes.

2.2.3 Predictions and Analysis of Novel Coronavirus

A study done by (Vaishya *et al.*, 2020) suggested that Artificial Intelligence can be applied to COVID-19 pandemic, which can help to analyze, precaution, and how to tackle COVID-19 pandemic both socially and economically for individuals and government. They have put out seven different applications of artificial intelligence. They are recognition of the infected individual, overseeing the treatment, contact tracing of the patients, forecasting of infected cases and mortality rate of the virus, researching on vaccines and drug medicines, relieving some overhead pressure on the healthcare professionals, and preventing the infection. They explained how AI could help in these areas and be helpful to everyone to get ready for the outbreak.

Another study focusing on the reproduction rate R_0 of 2019-nCov was carried out to see how the cases will increase in the future. The exponential growth of the infected cases is seen, and many countries are dealing with their own R_0 values. Li *et al.* analyzed and found the R_0 value can be at 4.04, 3.17, and 2.42, associated with 2-fold, 4-fold, and 8-fold, respectively (Zhao *et al.*, 2020).

A real-time virus transmission model was proposed by (Tan and Chen, 2020), which can predict quick short term COVID-19 infections transmissions. It uses Susceptible-Exposed-Infected-Removed (SEIR) virus transmission model as its base and adds some elements onto it. The study predicted that there would be peak infected persons in mid-July with 1.09 million cases with 90 thousand deaths. But it already failed the prediction and can be worked on to make the model better.

Another study used the Auto-Regressive Integrated Moving Average (ARIMA) and Exponential Smoothening (ES) to predict the cases in India. As both algorithms work on time-series data, the data is converted to time-series. The study concluded that although ARIMA worked better in the current scenario, the predicted result can be changed as the researchers did not have enough data to build a successful model (Gupta and Pal, 2020).

In other studies, Generalized Logistic Growth Model (GLM) was used to predict cases on infected persons in China. The model was evaluated on MSE, having a good-fit model with overtaking the accuracy of the other models (Roosa *et al.*, 2020). (Jia *et al.*, 2020) used three different models to predict the cases, Logistics model, Gompertz model, and Bertalanffy model. The methods of these models are the same, but mathematical models are different. The trend can be detected efficiently by the Logistic model, and the Gompertz model was better in fitting the data for non-Hubei province. The logistic model was the most fitting model than the rest of the models. Classifier model was used by (Al-Najjar and Al-Rousan, 2020), Neural Network was used to classify the dataset into multiple classes. This study reported the most important factor for predicting death was infection reason, infection date, and area. Sex and group are the least important factor in recovered cases. Selecting an important categorical variable with a numerical variable makes the model a better fit and accurate.

3. Methodology

In this study, CRISP-DM (Cross Industry Standard Process for Data Mining) methodology will be used for this research. CRISP-DM consists of six phases, namely, Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment (Fig 1). CRISP-DM reference model shows a life cycle of the project. It is a well-known and robust methodology and used all over the industry (Wirth and Hipp, 2000).

- Business Understanding: The first thing to see the objectives of the project from a business point of view and transforming knowledge into the problem statement.
- Data Understanding: Gathering datasets for the required problem, getting to know the dataset, and check the quality of the datasets.
- Data Preparation: Most of the ETL process (Extract, Transform, Load) to the required format.
- Modelling: Selecting and applying various models to get the best result out of it.
- Evaluation: Evaluate the model and check if the business requirements are met or not.
- Deployment: Deploying the project as required by the business.

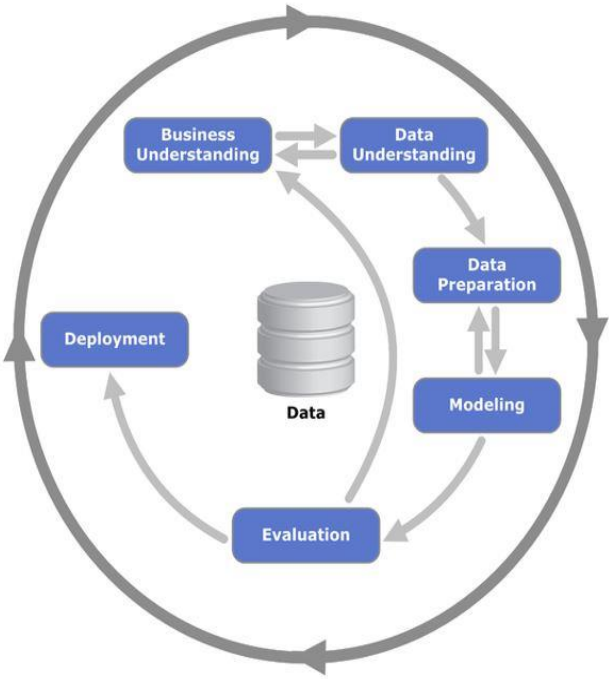

Fig. 1 Process Cycle of CRISP-DM

3.1 Architecture

The aim of this research is to predict the cases of infection of COVID-19 using machine learning algorithms.

- Raw data is loaded onto the notebook.
- Exploratory Data Analysis is done on the dataset.
- Different Machine Learning algorithm models will be used to fit the data.
- Predictions will be made with the help of models.
- All the models will be evaluated, and the best model will be chosen.

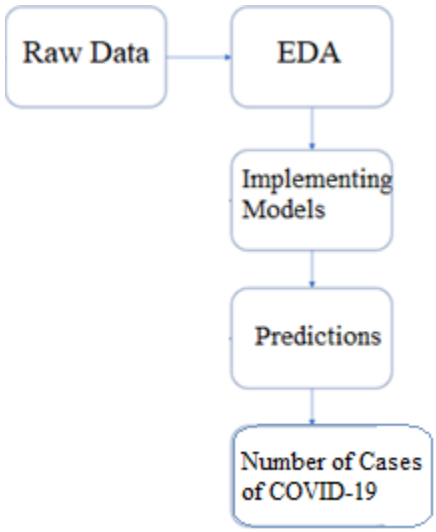

Fig. 2 Architectural Design of the Process

3.2 Data Gathering

The COVID-19 dataset is downloaded from the 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE; it is publicly available. Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL). It contains multiple CSV files and is regularly updating to keep it up to date. It can be used publicly for educational and academic research purposes. 3 CSV files are used in the research. Attributes are shown in Table 1.

Sr. No.	Data Type	Variable	Meaning
1	string	Province/State	States name
2	String	Country/Region	Country Name
3	num	Lat	Latitude
4	num	Lon	Longitude
5	dates	dates	Dates
6	num	Confirmed	Confirmed cases
7	num	Deaths	Deaths cases
8	num	Recovered	Recovered cases
9	num	Active	Active cases
10	date/time	Last Update	Last update time

Table 1 Data Description of COVID-19 Dataset

The data are contained in the files `time_series_covid19_confirmed_global.csv`, `time_series_covid19_deaths_global.csv`, `time_series_covid19_recovered_global.csv`, and `latest_date_updated.csv.` The data in the files have the above-mentioned attributes in the CSV file.

3.3 Methodology

For this research, three machine learning algorithms are used:

3.3.1 Support Vector Regression (SVR)

It is based on the Support Vector Machine (SVM) algorithm to forecast a continuous variable. Support vector regression tries to find a line that fits best under the same threshold value. It classifies all prediction lines in two types, one which passes via the error boundary line and the one which does not. The line which passes is considered a support vector for forecasted values (www.aionlinecourse.com, 2020).

$$Yi = (w, xi) + b \pm \acute{\varepsilon}$$

3.3.2 Polynomial Regression

Polynomial regression is a special case of linear regression where we fit a polynomial equation on the data with a curvilinear relationship between the target variable and the independent variables.

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \ldots + \theta_n x^n$$

$\boldsymbol{\theta 0}$ is the bias,
$\boldsymbol{\theta 1, \theta 2, \ldots, \theta n}$ are the weights in the equation of the polynomial regression,
and $\boldsymbol{n}$ is the degree of the polynomial ('Polynomial Regression | Polynomial Regression In Python', 2020).

3.3.3 Bayesian Ridge Regression

It estimates a probabilistic model of the regression problem, as described above. The prior for the coefficient w is given by a spherical Gaussian:

$$p(w|\lambda) = \mathcal{N}(w|0, \lambda^{-1}\mathbf{I}_p)$$

The priors over α and λ are chosen to be gamma distributions, the conjugate prior for the precision of the Gaussian. The resulting model is called Bayesian Ridge Regression and is similar to the classical Ridge (*1.1. Linear Models — scikit-learn 0.22.2 documentation*, 2018).

3.4 Performance Measure

3.4.1 Mean Absolute Error (MAE)

The Mean Absolute Error (or MAE) is the average of the absolute differences between predictions and actual values. It is a measurement of the regression model.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$ [1]

3.4.2 Mean Squared Error (MSE)

The Mean Squared Error (or MSE) is much like the mean absolute error in that it provides a gross idea of the magnitude of the error.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

3.4.3 Root Mean Squared Error (RMSE)

Taking the square root of the mean squared error converts the units back to the original units of the output variable and can be meaningful for description and presentation.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}.$$

3.4.4 R^2 Metric

The R^2 (or R Squared) metric provides an indication of the goodness of fit of a set of predictions to the actual values. In statistical literature, this measure is called the coefficient of determination.
This is a value between 0 and 1 for no-fit and perfect fit, respectively.

$$R^2 \equiv 1 - \frac{SS_{\mathrm{res}}}{SS_{\mathrm{tot}}}$$

3.5 Feature Engineering

Raw data is like crude oil which has to be refined to use the refined data in the machine learning algorithms (Koehrsen, 2018). The process of extracting features from the dataset is called as feature engineering. In this research, particular features are selected through the use of selecting required columns for analysis of data by the help of loading the file into a data frame and then selecting the columns in the data frame by the help of .loc[] functionality in python. Several columns like latitude, longitude, province, fips, admin2 and combined keys were left out while creating and choosing required fields.

4. Implementation

4.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) means to show the main characteristics of the datasets with the help of visual methods or simple statistical methods before any modeling.

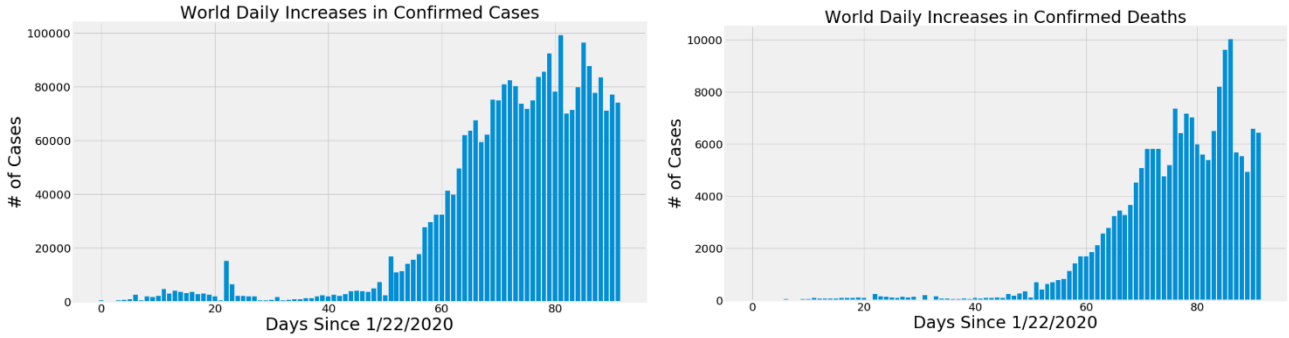

Fig 3 Bar plot of the daily world increase in confirmed cases and confirmed deaths

Fig 3 shows that there has been a sharp increase in confirmed cases in the world and almost reaching 100 K one day, and it shows the rate of deaths was slow at the beginning, then halfway through it started to get higher.

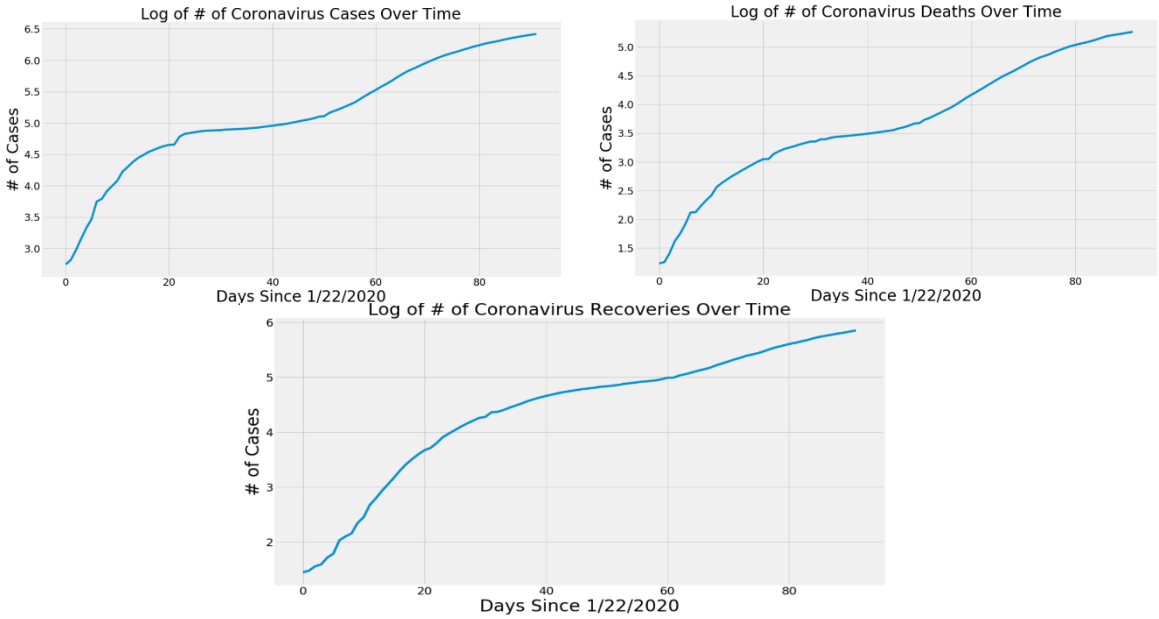

Fig 4 Logarithmic progression of cases

Fig 4 shows the logarithmic progression of cases, including infected, deaths, and recoveries. It shows a bit of dip midway but then quickly rises.

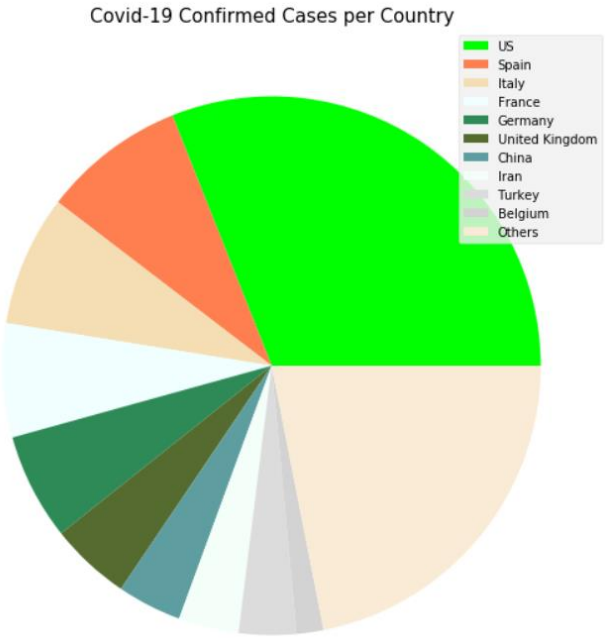

Fig 5 Pie chart of all countries

Fig 5 shows that the US has the highest number of cases all around the world, followed by Spain and Italy.

4.2 Models

4.2.1 Introduction

This section covers the implementation of different prediction machine algorithms. Methods are discussed, and important materials are also discussed. Python is used for this study, and several packages are imported. Jupyter Notebook is used for the environment.

4.2.2 Implementing different models for forecasting

For data analysis, Scikit-learn is one of the most powerful packages, which is the go-to package for any data analysis tasks in python. As this is regression, the author does not have to change categorical variables into the number variable. All the dates are converted into a specific date format. Then the dataset is split into train and test datasets. It is set to forecast the values for ten days in the future. Further down, all the implementation of the models are performed:

- Support Vector Regressor: It is called by using SVR() from sklearn.svm package. SVR kernel is set to poly and used RandomizedSearchCV() to find possible parameters for SVR, then svr.best_params_ is called to find and lock the best parameters for SVR. All the datasets are split into train and test sets by 80:20 ratio by using train_test_split().

- Polynomial Regression: For using polynomial regression, firstly, the degree of the data is transformed into a degree of 3. Then LinearRegression() is run on the training dataset. Train and test datasets were split into 80:20 ratio by using train_test_split()

- Bayesian Ridge Regression: For Bayesian ridge regression, the complexity of the degree was raised to be 4. RandomizedSearchCV() to find possible parameters for

bayesian_search, then bayesian_search.best_params_ is called to find and lock the best parameters for Bayesian ridge. All the datasets are split into train and test sets by 80:20 ratio by using train_test_split()

5. Evaluation

5.1 Evaluation of Support Vector Regressor

This regression model is evaluated on the basis of mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R^2 score. It is calculated by using SVR predicted value against the test dataset. MAE = 90745.4378, MSE = 15457583312.1607, RMSE = 124328.52795, and R^2 = 0.8273. (Fig 6)

```
MAE: 90745.437886391
MSE: 15457583312.160719
RMSE: 124328.52975950741
R2 Score:  0.8273301960487327
```

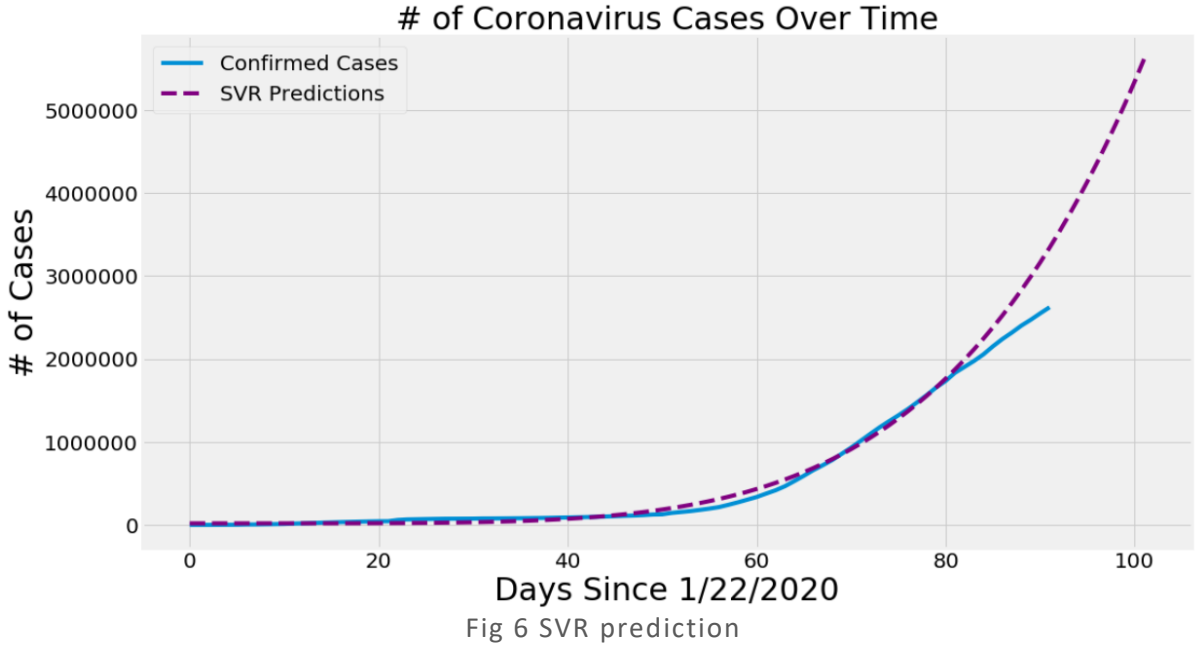

Fig 6 SVR prediction

5.2 Evaluation of Polynomial Regression

This model is evaluated based on mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R^2 score. It is calculated by using SVR predicted value against the test dataset. MAE = 60597.00667, MSE = 7055687107.96214, RMSE = 83998.13752, and R^2 = 0.90769. (Fig 7)

```
MAE: 60597.00667960047
MSE: 7055687107.962194
RMSE: 83998.13752674636
R2 Score:  0.9076913883959159
```

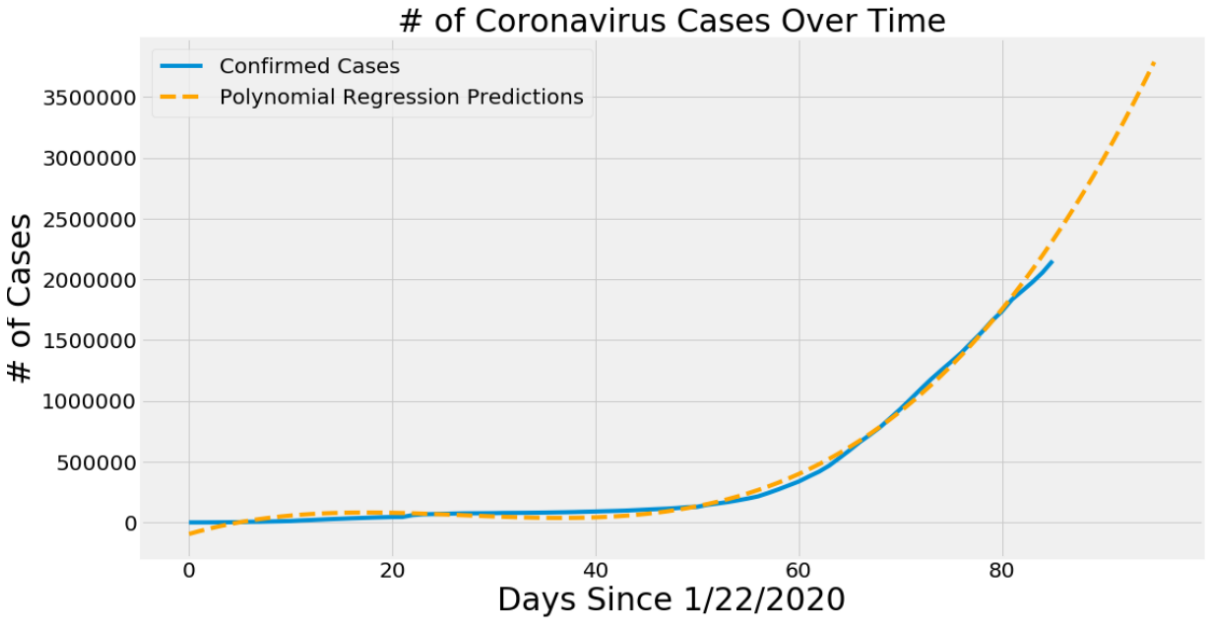

Fig7 Polynomial Regression Prediction

5.3 Evaluation of Bayesian Ridge Regression

This regression model is evaluated based on mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R^2 score. It is calculated by using SVR predicted value against the test dataset. MAE = 53659.525617, MSE = 5172591875.231913, RMSE = 71920.733277, and R^2 = 0.93213. (Fig 8)

```
MAE: 53659.52561778738
MSE: 5172591875.231913
RMSE: 71920.73327790751
R2 Score:  0.9321396159485351
```

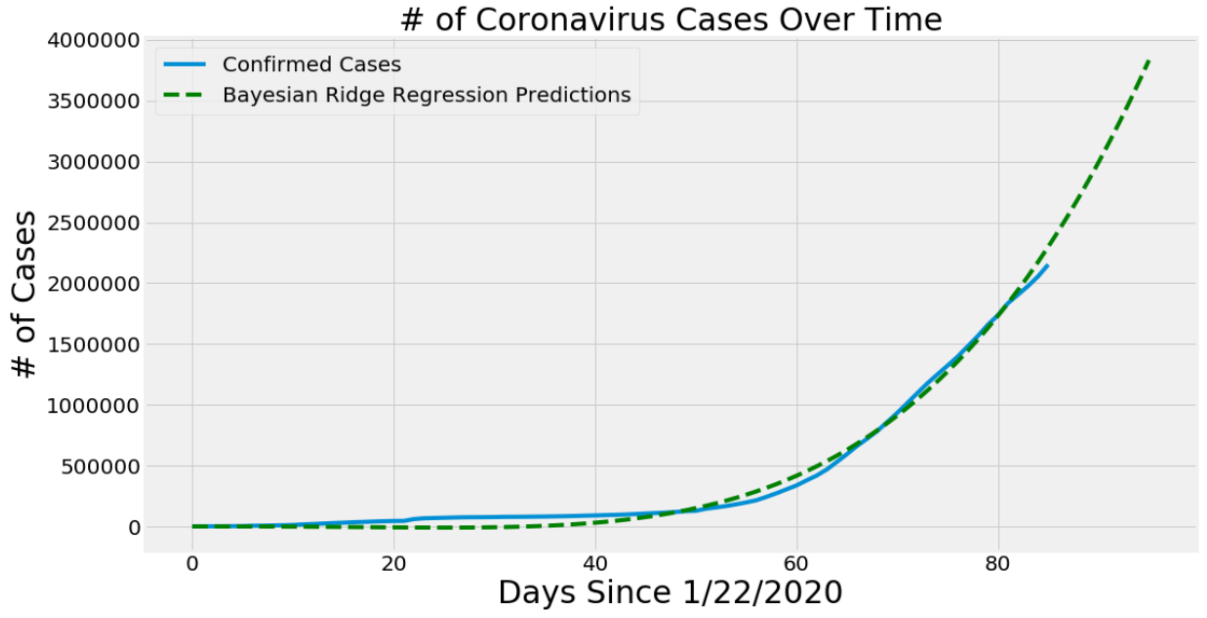

Fig 8 Bayesian Ridge Regression Prediction

5.4 Evaluation of Mortality Rate and Recovery Rate

Over the period of the months, the mortality rate has been fluctuating up and down. To take a better measure to take mean of the mortality rate with the help of NumPy. To ease any large jumps, the Mean Mortality rate now is y=0.0372331490, as shown below (Fig 9). The mortality rate is slowly increasing, which could prove harmful.

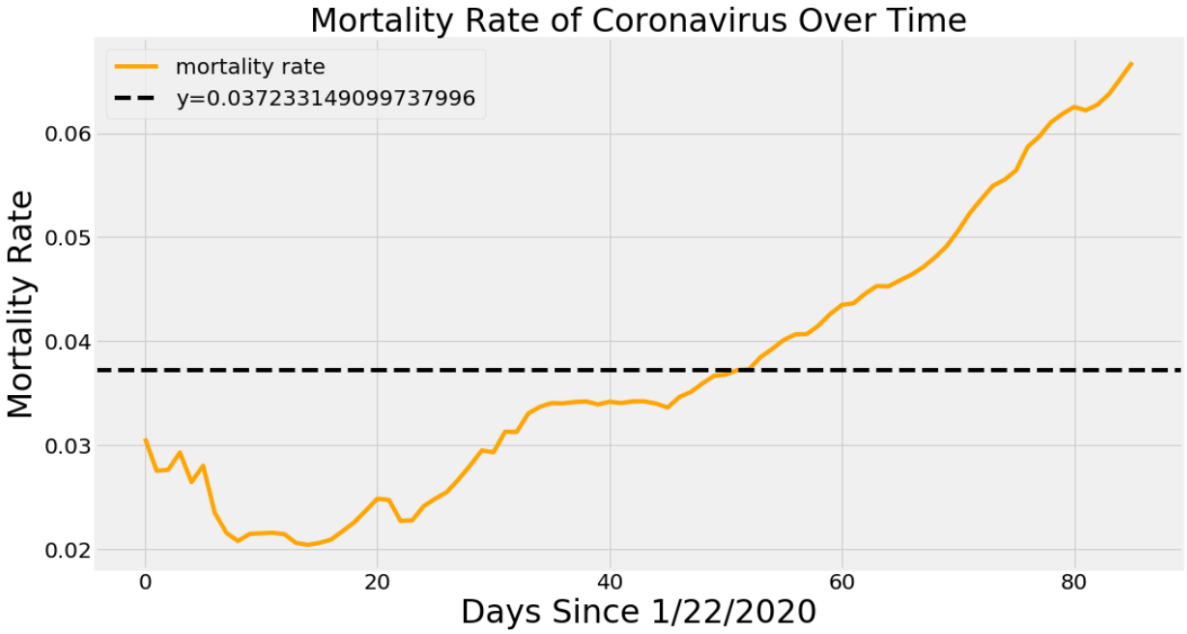

Fig 9 Mortality Rate over time

Like the mortality rate, the Recovery rate has also been changed over time. Similar to the mortality rate, the mean is taken of all the recovery rate. Mean Recovery rate now y=0.249286477, as shown below (Fig 10).

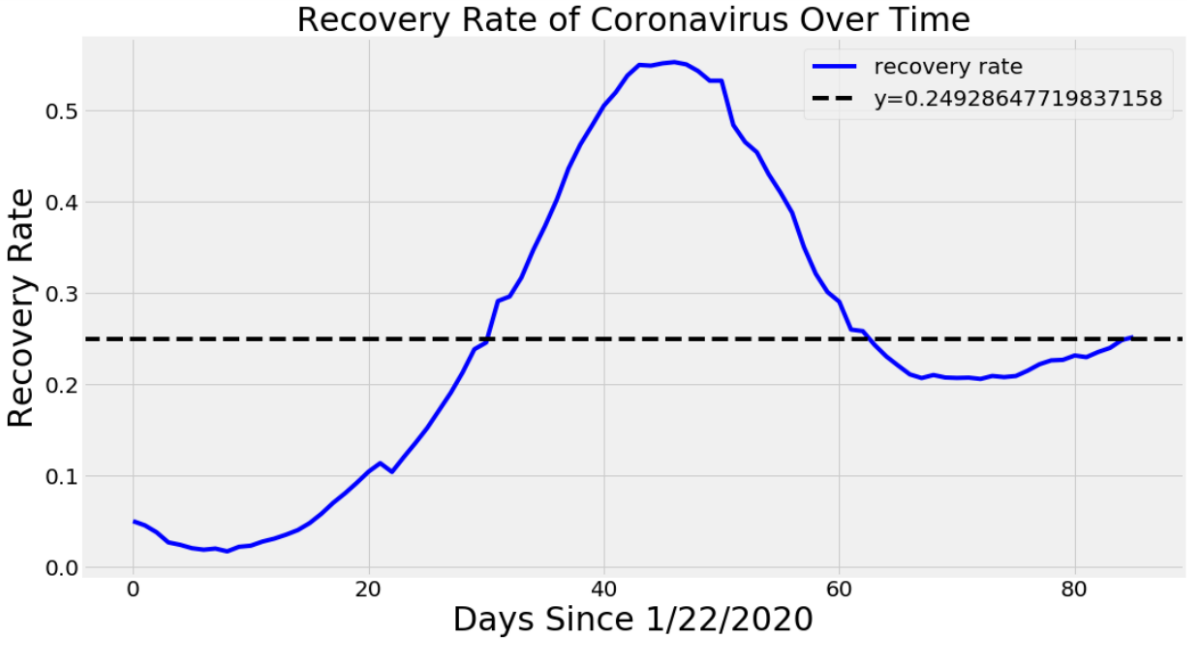

Fig 10 Recovery rate over time

5.5 Discussion

Forecasting any pandemic is a tedious process with many trial and error processes throughout the research process. As the dataset is ever-changing, and it is challenging to fit the models every time, and there are slim chances of never working out. There could be any pandemic next time, for this easily accessible data and information are equally necessary to start checking the need and understanding risks. This knowledge and knowing the consequences are needed to assess the risk and take control of the situation. As WHO announced on 22nd April 2020, COVID-19 is going to be in the society for the better part of the rest of the year. Hence, this study tried to give out some insights by analyzing real-time datasets and forecasting the probable cases to let people know how critical this virus is. As the dataset is evolving daily, Government measures imposing lockdowns, evolving of the virus itself could make a difference in the future predictions.

Part of the research, the author tried to incorporate many factors that would have to increase the model efficiency and accuracy. But due to inadequate resources and time crunch, they could not include these factors, such as the number of hospital beds, healthcare professionals, the number of hospitals, etc. This study will help to forecast the number of COVID-19 cases, mortality rates, and recovery rates all around the world.

Date	SVR Predictions	Polynomial Regression Predictions	Bayesian Ridge Regression Predictions
17/4/2020	2539482.0	2444875.0	2426935.0
18/4/2020	2689425.0	2575313.0	2560420.0
19/4/2020	2846422.0	2710287.0	2699234.0
20/4/2020	3010721.0	2849875.0	2843516.0
21/4/2020	3182572.0	2994153.0	2993406.0
22/4/2020	3362233.0	3143194.0	3149045.0
23/4/2020	3549967.0	3297076.0	3310576.0
24/4/2020	3746043.0	3455874.0	3478144.0
25/4/2020	3950736.0	3619663.0	3651897.0
26/4/2020	4164328.0	3785519.0	3831982.0

Table 2 Forecasted numbers of COVID-19 patients all around the world

	SVR	Polynomial Regression	Bayesian Ridge Regression
MAE	90745.437886391	60597.00667960047	53659.52561778738
MSE	15457583312.160719	7055687107.962194	5172591875.231913
RMSE	124328.52975950741	83998.13752674636	71920.73327790751
R^2 Score	0.8273301960487327	0.9076913883959159	0.9321396159485351

Table 3 Comparison between Prediction models

If you look at Table 3, you can see there is a comparison between prediction models based on the evaluation measure of regression like mean absolute error, mean squared error, root mean squared error and R^2 score. The lower the RMSE value is, the better the model performs. The R^2 score lies between 0 and 1, i.e., 0 being the no fit and 1 being the best fit model.

As you can see from Table 3, the Support Vector Regressor has the lowest R^2 score = 0.8273 among the three and the highest RMSE value = 124328.5297 amongst the three models, which tells us that the Support Vector Regressor is the least preferred model to choose from. While Bayesian Ridge Regression has the highest R^2 score = 0.9321 and the lowest RMSE value = 71920.7332. Hence, the best model in this study was found to be the Bayesian Ridge Regression model to predict COVID-19 cases all around the globe.

6. Conclusion and Future Work

This research study concludes that forecasting the COVID-19 pandemic can be done by using a regression model to predict future cases and take a better stand and be prepared for what is yet to come. Thus, from the analysis and evaluation of the Support Vector Regressor, Polynomial Regression, and Bayesian Ridge Regression individually, the Bayesian Ridge Regression model performed the best having a R^2 score of 0.9321 and RMSE value of 7190.7332 among other models.

Different studies have used a classification algorithm to predict the chances of dying of an infected person by its health, age, sex, and many other factors. This study can be extended by incorporating multiple factors like a number of hospital beds, level of the healthcare system and professional workers, area/region, lockdown measures, social distancing, government plans, etc. Also, researchers can look into previous medical records and can prevent the person from getting infected.

Acknowledgment

Throughout the research study, I have received numerous support and assistance. I would like to thank my supervisor, Mr. Jorge Basilio, whose expertise was essential in the construction of this research study.

References

1.1. Linear Models — scikit-learn 0.22.2 documentation (2018). Available at: https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression (Accessed: 23 April 2020).
Al-Najjar, H. and Al-Rousan, N. (2020) 'A classifier prediction model to predict the status of Coronavirus COVID-19 patients in South Korea', *European Review for Medical and Pharmacological Sciences*, 24(6), pp. 3400–3403. doi: 10.26355/eurrev_202003_20709.
Al-Turaiki, I., Alshahrani, M. and Almutairi, T. (2016) 'Building predictive models for MERS-CoV infections using data mining techniques', *Journal of Infection and Public Health*, 9(6), pp. 744–748. doi: 10.1016/j.jiph.2016.09.007.
Chowell, G. *et al.* (2020) 'Real-time forecasting of epidemic trajectories using computational dynamic ensembles', *Epidemics*, 30, p. 100379. doi: 10.1016/j.epidem.2019.100379.
Ding, Y. *et al.* (2020) 'Association between population migration and epidemic control of Coronavirus disease 2019', *Science China Life Sciences*. doi: 10.1007/s11427-020-1695-5.
Gupta, R. and Pal, S. K. (2020) *Trend Analysis and Forecasting of COVID-19 outbreak in India*. preprint. Public and Global Health. doi: 10.1101/2020.03.26.20044511.
Iannelli, M. and Pugliese, A. (2014) 'Mathematical modeling of epidemics', in Iannelli, M. and Pugliese, A., *An Introduction to Mathematical Population Dynamics*. Cham: Springer International Publishing, pp. 209–264. doi: 10.1007/978-3-319-03026-5_8.
Jia, L. *et al.* (2020) 'Prediction and analysis of Coronavirus Disease 2019', p. 19.

John, M. and Shaiba, H. (2019) ‘Main factors influencing recovery in MERS Co-V patients using machine learning’, *Journal of Infection and Public Health*, 12(5), pp. 700–704. doi: 10.1016/j.jiph.2019.03.020.
Koehrsen, W. (2018) *Feature Engineering: What Powers Machine Learning*, *Medium*. Available at: https://towardsdatascience.com/feature-engineering-what-powers-machine-learning-93ab191bcc2d (Accessed: 25 May 2020).
Madhav, N. *et al.* (2017) ‘Pandemics: Risks, Impacts, and Mitigation’, in Jamison, D. T. et al. (eds) *Disease Control Priorities: Improving Health and Reducing Poverty*. 3rd edn. Washington (DC): The International Bank for Reconstruction and Development / The World Bank. Available at: http://www.ncbi.nlm.nih.gov/books/NBK525302/ (Accessed: 22 April 2020).
‘Middle East respiratory syndrome’ (2020) *Wikipedia*. Available at: https://en.wikipedia.org/w/index.php?title=Middle_East_respiratory_syndrome&oldid=952427011 (Accessed: 22 April 2020).
‘Polynomial Regression | Polynomial Regression In Python’ (2020) *Analytics Vidhya*, 15 March. Available at: https://www.analyticsvidhya.com/blog/2020/03/polynomial-regression-python/ (Accessed: 23 April 2020).
Qiu, W. *et al.* (2017) ‘The Pandemic and its Impacts’, *Health, Culture and Society*, 9, pp. 1–11. doi: 10.5195/HCS.2017.221.
Raissi, M., Ramezani, N. and Seshaiyer, P. (2019) ‘On parameter estimation approaches for predicting disease transmission through optimization, deep learning and statistical inference methods’, *Letters in Biomathematics*, pp. 1–26. doi: 10.1080/23737867.2019.1676172.
Ray, E. L. and Reich, N. G. (2018) ‘Prediction of infectious disease epidemics via weighted density ensembles’, *PLOS Computational Biology*, 14(2), p. e1005910. doi: 10.1371/journal.pcbi.1005910.
Roosa, K. *et al.* (2020) ‘Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020’, *Infectious Disease Modelling*, 5, pp. 256–263. doi: 10.1016/j.idm.2020.02.002.
Sedaghat, A. *et al.* (2020) ‘COVID-19 protection guidelines in outpatient medical imaging centers’, *Academic Radiology*, p. S1076633220302129. doi: 10.1016/j.acra.2020.04.019.
Tan, S. X. D. and Chen, L. (2020) ‘Real-Time Differential Epidemic Analysis and Prediction for COVID-19 Pandemic’, *arXiv:2004.06888 [q-bio]*. Available at: http://arxiv.org/abs/2004.06888 (Accessed: 22 April 2020).
Tripathi, R., Reza, A. and Garg, D. (2019) ‘Prediction of the disease controllability in a complex network using machine learning algorithms’, p. 12.
Vaishya, R. *et al.* (2020) ‘Artificial Intelligence (AI) applications for COVID-19 pandemic’, *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4), pp. 337–339. doi: 10.1016/j.dsx.2020.04.012.
Wang, J.-F. *et al.* (2008) ‘Data-driven exploration of “spatial pattern-time process-driving forces” associations of SARS epidemic in Beijing, China’, *Journal of Public Health*, 30(3), pp. 234–244. doi: 10.1093/pubmed/fdn023.
Wirth, R. and Hipp, J. (2000) ‘CRISP-DM: Towards a Standard Process Model for Data Mining’, p. 11.
www.aionlinecourse.com (2020) *Support Vector Regression | Machine Learning | Artificial Intelligence Online Course, www.aionlinecourse.com*. Available at: https://www.aionlinecourse.com/tutorial/machine-learning/support-vector-regression (Accessed: 23 April 2020).
Zhang, J. *et al.* (2020) ‘Risk factors for disease severity, unimprovement, and mortality of COVID-19 patients in Wuhan, China’, *Clinical Microbiology and Infection*, p. S1198743X20302172. doi: 10.1016/j.cmi.2020.04.012.
Zhao, S. *et al.* (2020) ‘The basic reproduction number of novel coronavirus (2019-nCoV) estimation based on exponential growth in the early outbreak in China from 2019 to 2020: A

reply to Dhungana', *International Journal of Infectious Diseases*, p. S1201971220300837. doi: 10.1016/j.ijid.2020.02.025.